Paula Vasquez-Henriquez

Gender Differences in Transport Perception using Social Media Data

Master of Science in Engineering

Advisor: Eduardo Graells-Garrido, Ph.D
Co-advisor: Diego Caro, Ph.D
Universidad del Desarrollo
Faculty of Engineering
Data Science Institute

July, 2020

Dedicated to my imposters.

It is good to have an end to journey toward; but it is the journey that matters, in the end.

— Ursula K. Le Guin

ACKNOWLEDGEMENTS

First and foremost, I thank God. Through all of my life, the ”Pulento” continues to challenge me to become the person I’ve always wanted to be, despite my resistance.

I thank my advisor Dr. Eduardo Graells-Garrido in brief words, although for everything he has done for me, he deserves several full pages. Thank you for seeing something in me, for teaching me, supporting me, challenging me, containing me, forcing me to grow, showing me new ways of seeing the world, and also for the countless jokes, lessons, bus interviews, sarcasm, Icochan stickers, unexpected stories, and tears at Café 339.

I also thank my co-advisor, Dr. Diego Caro, for his patience, guidance, infinite support, and teachings beyond academics. I treasure all the conversations we had, and appreciate your help keeping my feet on the ground.

I appreciate the contributions of many people who supported me during this process. To the members of the Institute of Data Science: Dr. Loreto Bravo, Dr. Leo Ferres, Dr. Daniela Opitz and Francisca Varela, who in one way or another contributed with support and teachings. Special thanks to Daniela, who helped write this defense. Thank you Dani for teaching me about writing!.

I want to thank my colleagues on this journey: Yerka Freire, Ignacio Pérez and Fernanda Sánchez. Sharing the joys and frustrations of this was crucial, and you made it far less lonely. Along the same lines, I want to thank Daniela Bonacic, who supported me from the first moment that I decided to venture into this city. Thank you very much Dani for walking with us on this adventure!.

Of course, I thank my mom and dad: Angélica and David. If I am in this place, it is thanks to the effort, dedication, hard work, and infinite love they have given me during all these years. Thank you for always believing in me and being my biggest fans.

I thank my sisters Violeta, Gabriela, Francisca, Isidora, Verónica and my brother Camilo, who were always there to support me, hug me and make me feel at home even while being away. Particularly, my sister Pilar, my rock. There are no words to thank you for being with me through thick and thin, since we were little, and forever.

To my love, Camilo: physical distance was not and will never be an impediment for our relationship. Thank you for the most beautiful lesson in the past couple of years: tender and patient love.

To my friend Belén: for being a sister, a roomie and a partner in suffering.

I want to give special thanks to Icochan, who was ever present in this journey.

Last but not least, I would like to thank the Faculty of Engineering for supporting my studies with a scholarship. Also, I would like to thank Dr. Felipe Baesler, who supported me during my first year.

Paula

This thesis was partially funded by CONICYT Fondecyt de Iniciación project #11180913 and by Concurso Interno de Investigación UDD #CI18.

RESUMEN

Las personas a menudo basan sus decisiones de movilidad en aspectos subjetivos de la experiencia de viaje, como la percepción del tiempo, el uso del espacio y la seguridad. Es bien sabido que diferentes grupos dentro de una población reaccionarán de manera diferente al mismo viaje, sin embargo, los métodos actuales de recopilación de datos podrían no considerar los aspectos multidimensionales de la percepción del viaje, lo que podría llevar a pasar por alto las necesidades de grandes grupos de población. En este trabajo, proponemos medir varios aspectos de la experiencia de viaje desde Twitter, con un enfoque en las diferencias con respecto al género. Analizamos más de 400,000 tweets de 100,000 usuarios sobre el transporte en Santiago, Chile. Nuestros principales hallazgos muestran que ambos sexos se expresan de manera diferente, las mujeres escriben sobre sus emociones con respecto al viaje (tanto sentimientos positivos como negativos), y los hombres se expresan usando jerga, lo que dificulta la interpretación de las emociones. La diferencia más fuerte está relacionada con el acoso, no solo en el transporte, sino también en el espacio público. Dado que estos aspectos generalmente se omiten de las encuestas de viaje, nuestro trabajo proporciona evidencia de cómo Twitter permite la medición de aspectos del sistema de transporte en una ciudad que se han estudiado en términos cualitativos, complementando las encuestas con aspectos emocionales y de seguridad que son tan relevantes como aquellos medidos tradicionalmente.

ABSTRACT

People often base their mobility decisions on subjective aspects of travel experience, such as time perception, space usage, and safety. It is well recognized that different groups within a population will react differently to the same trip, however, current data collection methods might not consider the multi dimensional aspects of travel perception, which could lead to overlooking the needs of large population groups. In this paper, we propose to measure several aspects of the travel experience from the social media platform Twitter, with a focus on differences with respect to gender. We analyzed more than 400,000 tweets from 100,000 users about transportation from Santiago, Chile. Our main findings show that both genders express themselves differently, as women write about their emotions regarding travel (both, positive and negative feelings), that men express themselves using slang, making it difficult to interpret emotion. The strongest difference is related to harassment, not only on transportation, but also on the public space. Since these aspects are usually omitted from travel surveys, our work provides evidence on how Twitter allows the measurement of aspects of the transportation system in a city that have been studied in qualitative

terms, complementing surveys with emotional and safety aspects that are as relevant as those traditionally measured.

CONTENTS

1 INTRODUCTION

Transport experience varies from different groups within a population [1]. Men and women have different travel characteristics [2]–[4], place different degrees of importance on mobility-decision factors (*e.g.* time, cost, accessibility, frequency, safety), and move differently across the city depending on their daily activities [5]–[8]. This difference is critical for the sustainability of cities, as stated in one of the Sustainable Development Goals by the United Nations, which aims to "provide access to safe, affordable, accessible and sustainable transport systems for *all*, [...] with special attention to the needs of those in vulnerable situations, *women* [...]" (emphasis ours).[1]

Advancing toward reaching this goal is not an easy task. One of the main difficulties is the quantifying of the differences in population groups. It is vital for transport planners to take into consideration the trends and behaviors of different segments of the population and understand their context (conditions, opportunities, and constraints) [9]. Traditional methods to gather information on travel experience often observe users as one "average user" group, without considering the heterogeneity among groups of people. For instance, surveys are the most common tool to measure the travel experience [10], [11]. Typically, their design is targeted to optimize quantitative characteristics such as costs for the system, while maintaining or improving quality of service, mainly, time and frequency. Even though they offer rich information, surveys have limitations. Surveys are not created to consider the multi-dimensional aspects of perception when people travel [12], leading to conclusions assuming that the travel experience is the same for both genders [13]. Reportedly, that is not the case. For instance, women experience higher levels of violence and harassment in transportation, which leads them to respond in ways that affect their travel experience and quality of life [14]. Many of these incidents are under-reported to the authorities, thus lowering the probability of this problem being accurately reflected in urban and transport planning [15]. The reasons include entrenched gendered power hierarchies and the lack of robust information about women's needs.

The ubiquity of mobile phones has allowed people to express their opinions and daily experiences on social media platforms, creating a vast source of unstructured data on implicit user satisfaction. We propose that social media data may complement surveys and other methods. In particular, in this study, we analyze gender differences in transport perception through the analysis of posts on Twitter. We hypothesize that, if gender influences travel experience in any way, this should be reflected by a difference in the linguistic components of the texts published in social media. Measuring the perception and its potential gender differences is a challenging task due to the unstructured data and the informal nature of social media.

1 https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-11-sustainable-cities-and-communities.html

The research question that drives our study is: *How can we measure the gender differences in transport perception presented in social media?*. To answer this question, we develop a three-step process to analyze transport-related opinions expressed on social media platforms. First, we infer gender for user profiles, which is not directly available on micro-blogging sites. Second, we extract the latent structure from the discussion using a topic modeling approach [16], [17], which allows to separate the personal experience (e.g., feelings about a specific trip in metro) from the general discussion (e.g., opinions on public transport policy). Finally, we quantify perception and its differences with respect to gender with an analysis grounded on psycho-linguistic theory, which allows us to study the affective and relative aspects of transportation revealed by users.

With this methodology, we study the Twitter transport perception in Santiago (Chile) during 2017-2018. We chose Chile as it is the most connected country in Latin America since 2017 [18], and Santiago as it is the largest city in Chile, with a population of 7 million people. Santiago also poses a wide offer of transportation modes, including metro, bus, taxi, shared taxi, public bikes, and several ride-share apps for cars, scooters, and bikes. We chose Twitter as it is one of the most accessed applications from mobile phones in Santiago [19].

The main contributions of this work are two-fold. The first contribution is a methodology to analyze the discussion about transportation in social media, with a focus on gender differences. This contribution could help data scientists to extract new applied insights from already available data. The second contribution is a case study of measured gender differences, using Twitter data from a big city. These contributions could help urban planners to widen their understanding of how different groups of people experience one of the most recurrent daily activities, transportation, using a promising data source in planning for sustainability: social media [20].

This work produced a conference paper, which can be found at the Appendix A, and a paper in a journal indexed in Web of Science. Both citations are provided below:

1. Vasquez-Henriquez, P., Graells-Garrido, E., & Caro, D. (2019, June). Characterizing Transport Perception using Social Media: Differences in Mode and Gender. In *Proceedings of the 11th ACM Conference on Web Science* (pp. 295-299).

2. Vasquez-Henriquez, P., Graells-Garrido, E., & Caro, D. (2020). Tweets on the Go: Gender Differences in Transport Perception and Its Discussion on Social Media. *Sustainability*, 12(13), 5405.

Additionally, different participations were carried out in conferences and other activities. In 2018 we participated in the Y4PT Hackathon with the project "Mochi," which aimed to use emojis as proxies for transport perception and received an honorable mention award. With a follow-up of that project, we participated in the Innovation Fair of Universidad del Desarrollo, for which we received 1st prize in presentations. That year we also presented our on-going work at Universidad de Chile's "Encuentro de Urbanistas del Mañana," and also gave a talk at the StarsConf 2019 conference. In 2019 we presented our work at the NetSciX Santiago conference, which caught the attention of the Inter American Bank of Development who invited us to be keynote speakers at their Gender and Transportation seminar in Washington D.C. That same year we presented a poster at NetMob 2019. In 2020 we had our abstract accepted at the 1st New Technologies for Urban Mobility workshop.

This dissertation is structured as follows. Chapter 2 presents the literature review on social media analysis and transportation experience measurement. Chapter 3 introduces the data and the proposed methodology for measuring gender differences in transport perception using social media data. Chapter 4 shows the results for a case study in Santiago, Chile. Chapter 5 presents the implications and main limitations of this work, as well as future work, and final conclusions of this dissertation.

2 | BACKGROUND

There are two main areas of research related to this work: *social media analysis* and *transportation experience measurement*. Figure 1 shows a diagram of these two areas, which we will discuss in relation to the contributions of this work.

2.1 SOCIAL MEDIA ANALYSIS

In the field of Social Media Analysis (SMA), we identify three sub-areas relevant for this work: *demographic analysis*, *sentiment analysis*, and *topic modeling*.

2.1.1 Demographic Analysis

Demographic analysis is essential, as social media platforms are not representative of the entire population in terms of age, gender, and socio-economic status [21]. Data obtained from this type of source cannot be generalized, although it is possible to de-bias the results from the analysis by using census data and specific methods [22]. In our work, we put the focus on gender, a specific demographic attribute of user profiles. Since this attribute is not always explicit, researchers need to infer them using machine learning methods [23]. Then, we analyze each group in separate, meaning that even if there are imbalances in the analyzed population groups, the insights regarding each gender should not be affected by this imbalance.

2.1.2 Sentiment Analysis

Sentiment Analysis is the measurement of polarity in text, mainly between negative, neutral, and positive [24]. Methods to extract polarity commonly start with a subjectivity lexicon of words that are associated with positive and negative feelings [25], which are then used to train a classifier [26]. Note, however, that the sentiment in a text may go beyond being positive or negative. Other approaches take into account the multidimensional aspects of speech. To the extent of our knowledge, the most common is Linguistic Inquiry and Word Count (LIWC) [27], which includes multiple emotions and other topical dimensions of expression such as psychological and cognitive components. LIWC usage with social media data includes predicting postpartum changes in emotion [28], identifying health issues [29], [30], and predicting political sentiment [31]. Our work is based on LIWC as a proxy of perception in transport (instead of other issues or themes), as

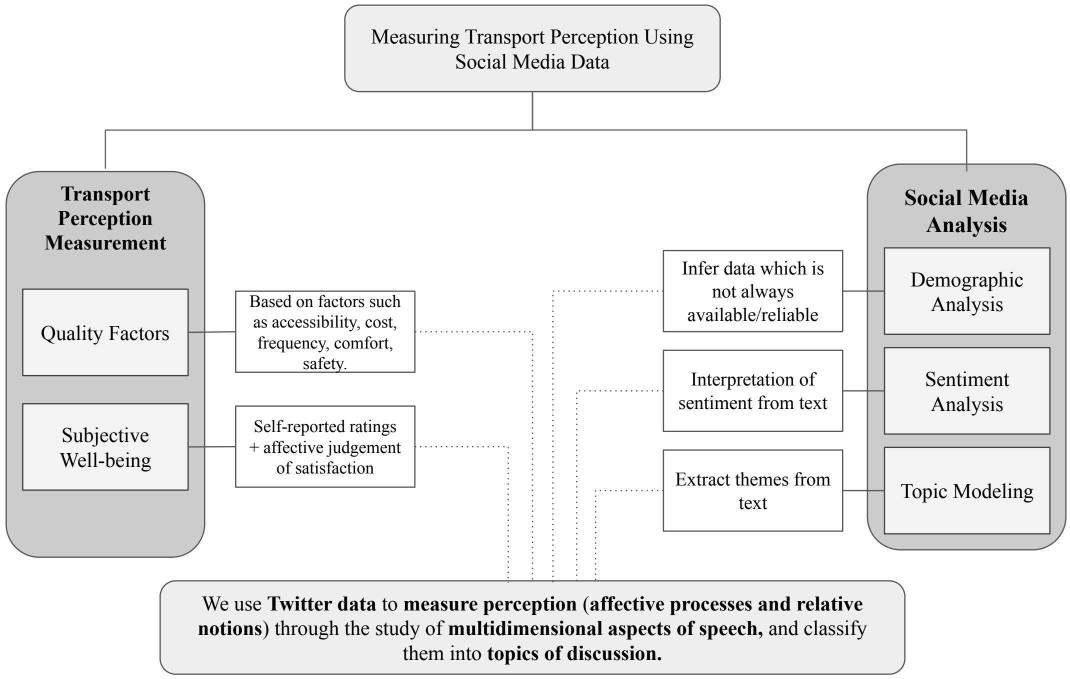

Figure 1: Conceptual model of the areas of research covered in this work.

it is available in Spanish (the language we analyze) and it has a proven track record in analyzing perception in different issues.

2.1.3 Topic Modeling

Topic Modeling is a family of approaches to extract thematic clusters from unstructured collections of text. In the case of social media, this refers to inferring the several themes, or *topics*, that people discuss. The most well-known methods are Latent Dirichlet Allocation (LDA) [32] and Non-Negative Matrix Factorization (NMF) [33]. Both methods infer latent topics in text, where words and documents are associated with these topics in a quantifiable way. LDA follows a generative approach, aiming at finding the most likely word and topic distribution given the text under analysis. NMF decomposes documents (or arbitrary objects represented as a non-negative matrix) into the sum of its latent parts. When choosing one method over the other, the task to be solved needs to be considered as well as the type of text. In short texts (which are commonly found in Twitter), NMF and its variations tend to exhibit topics of better quality [34]. In similar contexts to ours, NMF has been used to characterize urban areas according to their tweets [35], to infer political leaning of words and user profiles [36], and to separate the several topics related to depression discussion [37], among others. A potential limitation in this area is that latent topics are not always interpretable. To guide the topic extraction toward a predefined set of human-interpretable topics, a semi-supervised method can be used [38]. In topic modeling, these methods require modelers to label a fraction of the dataset, where labels may be applied to documents (such as Labeled-LDA [39]) or either to documents or keywords, such as in Topic-Supervised NMF (TS-NMF) [16]. Here we use TS-NMF due to its direct formulation, as well as its precedents in being used jointly with LIWC to characterize perception in at least two thematic discussions: migration [40] and mode of transportation usage [41].

2.2 TRANSPORTATION PERCEPTION MEASUREMENT

The concept of transport perception can be defined as the social value of the transport experience, while transport service quality can be considered as the sum of the users perceptions of that experience. In order to guarantee high quality transport service for users and to provide equal access of opportunities for all groups of society, understanding users transport perception is critical.

Transport perception is generally estimated focusing on quality related factors such as *accessibility, cost, frequency, reliability, comfort* and *safety* or through the measurement of subjective well-being. Although the mobility decisions of users of public transportation are assumed to maximize utility and satisfaction [42], transport perception has been proven to directly influence future user behaviour [43].

Some studies involving quality related factors include the creation and application of surveys which explicitly ask to participants to rank factor importances [44]–[49], and the generation of models to measure the the overall satisfaction considering factor contributions [50]–[55]. Many studies have attempted to identify peculiar characteristics of subpopulation groups on travel perception and behavior. These include defining traveler profiles based on the disaggregation of behavioral, socio-economic and psychosocial variables that map common attitudes, perceptions and experiences [56]; the differentiation of different periods of the day (holiday or non-holiday) and visitor profile (tourist or non-tourist) [57]; the usage of mode of transportation also shapes experience, as socio-economic characteristics alone do not explain travel patterns [58] and finally, there is a potential of mutual influence between attitudes and behavior over time [59].

The aforementioned works imply that analyzing the experience and its perception is no easy task. However, methods to measure perception are arguably standardized. The study of transport perception through the measurement of subjective well-being is based on Satisfaction with Travel Scale (STS) [60]. This scale considers both cognitive judgment (self-reported ratings) and affective judgment of satisfaction (duration and intensity of affects during a given time span) to measure perception. For example, Gatersben and Uzzel [61], compare different modes of transportation finding that walking and cycling are associated with higher levels of arousal. They also find that public transportation had a more negative transport perception given the lower levels of arousal due to delays transport and user waiting times. Other study developed by Eriksson et al. [62], comparing modes of transportation by the use of attributes such as amusement, finds that users perceive the activity "driving" as more fun and flexible than the activities related to the use of public transportation. A study on travel and time use on subjective well-being proved accessibility in transportation as a key factor towards improving the quality of life of women [63]. This furthermore highlights how transportation analysis should consider the particular gender barriers in the access to the city [64].

2.2.1 Gender Differences in Transport Perception

The relevance of the different aspects involved in transport perception varies by group based on their needs, patterns and the roles they play in society. For the case of women, is generally known that women have different travel patterns than men, as both social characteristics and environmental factors have been proven to influence their transport perception and travel behavior.

Studies show that women commute shorter distances [65]–[68], engage in more non-work trips related to house chores [69], [70], have a stronger tendency to link trips [4], rely more on public transportation or walking while having unequal access to cars [66], [71], and are more likely to respond to changing travel circumstances than men.

Studies on woman transport perception have been mainly addressed considering women's perception of safety, which often refers to the fear of sexual victimization [72]. There are plenty evidence supporting that emotions such as fear and anxiety are important detractors for women's mobility, which can make them avoid routes perceived as unsafe, only use them during daytime, refrain from using certain modes of transportation, which eventually restrains their mobility altogether [5], [7], [73]–[76].

Although there is extensive research on understanding women's travel characteristics and their fear of victimization in transportation, there has not been a lot of research concerned on understanding their general perception over quality factors. One of the main limitations to perform such studies is that the data on which these studies rely on, are usually collected through customer satisfaction surveys (on-board, online, phone or focus groups).

Even though of the aforementioned limitation, there a few studies that have reveled interesting results. Rojo et al. [77] performed the first study which assesses gender differences in public transport perception, showing that generally, women place higher value on punctuality, comfort, accessibility and safety when making their transport decisions. Other study done by Fan et al. [78] further elaborates on how waiting times in transportation were perceived by women, especially in spaces that were considered unsafe.

2.2.2 Social Media on Transport Perception

Social media gives their users the opportunity to express their opinions and tell experiences freely and spontaneously, creating a vast source of unstructured data.

Figure 2: Sample of a Twitter profile (Metro de Santiago/Santiago Subway)

Twitter, a popular micro-blogging platform that allows people to publish and exchange posts up to 280 characters (tweets), is extensively used for share feelings and point of views about a diversity of topics in real time. It has been shown that the range of topics discussed on Twitter are similar to those covered by news media [79].

Twitter's audience is huge, reporting by 2019 an average of 330 million monthly active users, which includes scientists, celebrities, company representatives and politicians, among others. Tweets generated by Twitter users may contain mentions to other users, hashtags to indicate themes within the post, URLs, emoji, *etc.*. A sample of a Twitter profile is shown on Figure 2.

As the twitter audience and topic posted on the platform are diverse, it is possible to collect text posts of users from different social and interests groups, that have motivated to researchers, companies and social organizations to use these platforms to get a sense of general sentiment of a specific topic such as the perception of immigrants, products or public and private services.

The role of social media in transport analysis has rapidly grown over the last years, allowing the ability to obtain information regarding trips [80] and activities [81], while highlighting the benefits of using these unstructured data sources [82]. There are few approaches to understand transport perception using this data source: in our literature search, we found a descriptive analysis of public transport perception from tweets in the city of Chicago [83], the monitoring of malfunctions in public transport in Madrid [84], the analysis of satisfaction with public transport in Santiago [85], and the public transport opinion (in terms of polarity, from negative to positive) in Nanjing [86].

Our work can be seen as a deeper analysis of the travel experience than these previous works, which are primarily focused on the polarity of messages, without considering gender differences nor more nuanced perception categories. This awareness of travel experience differences could allow transportation planners to consider a wider range of needs and dynamics into their work, beyond the needs of the business, complementing traditional data sources with fine-grained data.

3 DATA AND METHODOLOGY

In this chapter, we describe the dataset and methodology used to answer our research question. The context is the discussion about transportation from the city of Santiago, Chile. This discussion is contained in a collection of messages from the micro-blogging platform Twitter.

3.1 SOCIAL MEDIA DATASET

Twitter is a micro-blogging platform that allows people to publish and exchange posts up to 280 characters, called *tweets*. Tweets may contain mentions to other users, hashtags to indicate topics of the post, website addresses, emoji, *etc.*. Given that it is one of the most accessed applications from mobile phones in Santiago [19], we expect that people will frequently report their daily experiences, including transportation. Thus, we collected tweets related to transportation from the years 2017 and 2018. We used a manually crafted dictionary of keywords to query the Twitter Streaming API (Application Programming Interface). The list contained transportation-related words, hashtags, mentions of transportation accounts, URLs, and transportation emojis. Samples of these query terms can be found in Table 1.

In total, we collected 443K tweets from 112K users living in the Santiago Metropolitan area. A manual inspection of these tweets showed several themes within transportation, for instance, complaints mentioning the subway service provider, announcements of new bicycle paths, and general complaints about the size of the children safety chairs in cars (see Table 2 for examples).

The weekly publishing rate of those tweets is not uniform in time; on the contrary, people react to certain events. The periods with most volume have identifiable events related to transportation, however, not necessarily to the transportation experience. Topics include public transport service shutdowns and malfunctions and massive power outages, among others (see Figure 3). Likewise, the daily frequency also changes from day to day, as more tweets are published during business days than on weekends (see Figure 4). The hourly frequency distribution exhibits morning and evening peaks related to commuting times in Santiago (see Figure 5). Conversely, it does not resemble the general usage of the application in the city [19]. For instance, during lunch hours, people tweet less about transportation.

With this data set, we ought to answer our research question on understanding gender differences in transport perception. Next, we explain our methodology. It is composed of three main steps (see Figure 6 for a schematic diagram): (1) user profiling used to infer gender, (2) inference of transportation topics in the discussion, and (3) measurement of perception with gender differences.

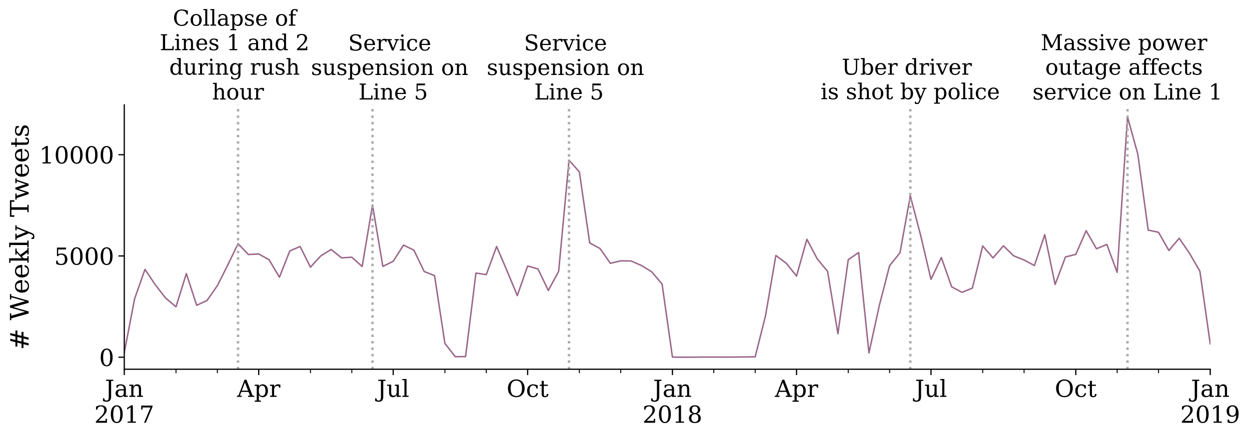

Figure 3: Transportation tweeting frequency during the years 2017 and 2018. Periods with zero tweets correspond to shutdowns in the tweet collection program.

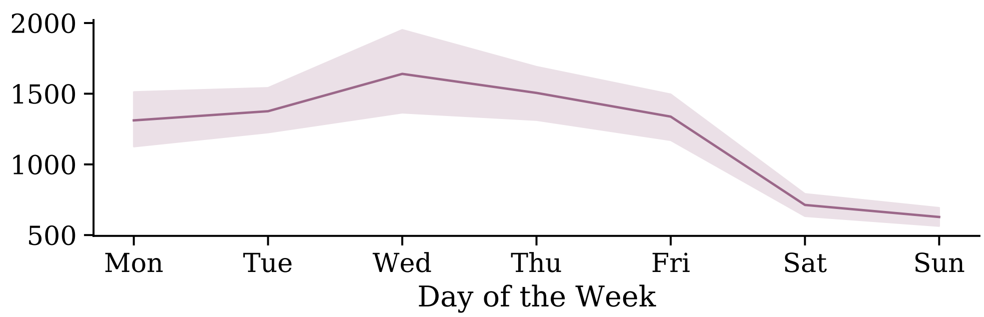

Figure 4: Average number of tweets per day of the week with a 95% confidence interval.

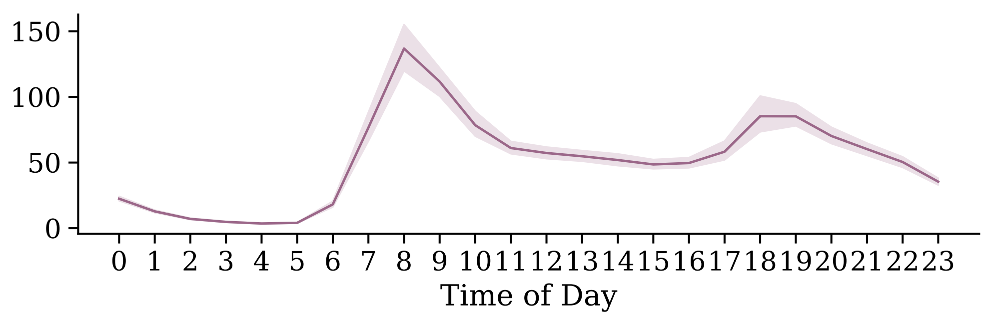

Figure 5: Average number of tweets per time of day with a 95% confidence interval.

Table 1: Samples of keywords used to crawl transportation tweets. In parentheses, translations or explanations.

Sample keywords		
bus stop (*paradero*)	easy taxi (ride-hailing application)	gas (*bencina*)
ride bicycle (*andar en bicicleta*)	panne (to run out of gas)	fare meter (*taxímetro*)
@uber_chile (ride-hailing application)	@mfc_oficial (bike organization)	🚌
@cabify (ride-hailing application)	train (*tren*)	🚗
bus-only lane (*corredor de buses*)	@metbus (bus operator)	#baquedano (*metro station*)
waze (GPS, route application for cars)	bike (*bici*)	subway station (*estación de metro*)
@autop_central (highway operator)	wagon (*vagón*)	🚒
highway (*autopista*)	🚴	#electricalbuses (*#buseselectricos*)
#publictransport (*#transportepublico*)	bus	🚕

3.2 USER PROFILING

In Twitter, user profiles contain the following attributes used in our study: an *id* (a number unique for each user), a *full name* (freely reported), a *description* (usually used as biography), a *location* (freely reported), and the number of *tweets*. Although Twitter requests such information, these fields are not mandatory, and can be filled with fictional data. For instance, a certain user could decide to write a fictional place (*e.g.*, *Wonderland*) as their profile location [87], or use a fantasy name, or no name at all. In this step of the methodology, we infer gender (binary) using the available user attributes and published content by using information provided by some users –who we can identify as from Santiago, with a specific gender–, to predict the attributes of others.

To predict these attributes XGBoost [88], a gradient boosting classifier based on decision trees. Our approach is based on the idea of using profiles that self-report their attributes to train a classifier, and then to propagate attribute labels to the rest of the data set [23]. In order to classify these attributes, we characterize each user through a set of features, which include:

1. The content published in his/her/their tweets, separated into *terms*, where a term may be a word, a hashtag, an emoji, an URL, or an n-gram (*i.e.*, n single terms that appear consecutively). Not all terms appear on the feature matrix, as there are many terms that are not relevant for the task at hand, such as lexical errors. Hence, we discard terms that are used by less than 50 users (determined through manual exploration).

Table 2: Sample tweets from our data set. The original tweets are in Spanish, we have translated them here for clarity.

#	**Tweet**
1	@metrodesantiago An escalator breaks and the whole line is down? Over an hour from Macul to Tobalaba 😡😡😡
2	Recognize and locate the new bike lanes that @Muni_Stgo is implementing. Plan your mobility this #march
3	In conclusion, if you have 3 children under the age of 9, you must change your car for a 3 row van!! #semana24
4	@metrodesantiago the smell of tobalaba station is very toxic, please put masks on the poor people who work there all day 😷
5	There are way too many uber drivers who don't know Santiago 🤦‍♀️
6	@sebastianpinera @metrodesantiago @GloriaHutt @MTTChile @J_LDominguez @louisdegrange And the humiliation of subway drivers? Any public apology? Very unfortunate comment @sebastianpinera minimizing the work of the drivers of @metrodesantiago ... Do you do your job well? To a part of the population, not less, it also seems little. Remember it.
7	@mbachelet L3 I look forward to it 🤗
8	@Transantiago thanks!!! :)
9	@UOCT_RM How are the traffic lights now?, I have to go to the O'Higgins park from Quinta normal
10	@Transantiago bus line 223 took 40 minutes to get to pb114 stop 😡 this always happens in the evenings! terrible
11	All traffic returned to Santiago of Chile.
12	Las Condes will saction street harassment @TroncoTorrealba I hope Vitacura will also implement this! I say it because for months I walked scared across Bicentenary Park (part of my travel route) where I was harassed by a man who parked cars.
13	@valderrax Do you think street harassment does not exist? Eg I was riding a bicycle and a guy told me: I would like to be the saddle to have my face in your ass. Is that an unfortunate phrase? and isn't it harassment? Doesn't it affect my freedom?
14	20 minutes between my office's door to the exit of National Stadium station. I LOVE YOU LINE 6 OF @metrodesantiago 😍
15	Again buses 425 won't stop after 00:00hrs. What's going on? Don't you oversee? I'm angry. @Transantiago @DTPMet @AlsaciaExpress

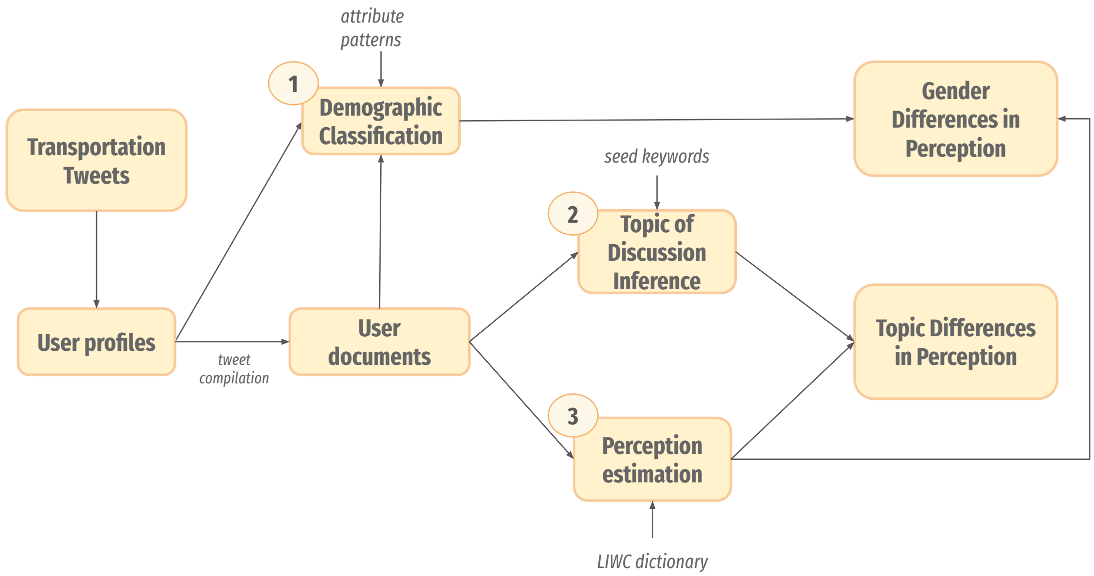

Figure 6: Schematic diagram of the proposed methodology.

2. The content published in his/her/their biographies, separated into terms (in the same way as tweet content). Here, we use a minimum threshold of 5 users for each term (determined through manual exploration).

3. The profile information, which includes time-zone, the home page link, and the usage of specific words related to commonly discussed topics (sports, religion, TV, etc.).

4. The interactions with other profiles through replying, retweeting, and quoting other tweets.

The idea is that using these features for users, a classifier may learn that specific language cues, interactions, words in profile description, among other characteristics, have predictive power.

Next, we identify which users self-report demographic attributes in their profiles. For instance, gender may be revealed by displaying a typically gendered name in the profile [89], whereas location may be disclosed by mentioning where a user lives in the profile [87]. The set of users that we can label though these simple measures comprise the training sets for the XGBoost classifiers, one for gender and one for location. Note that we use the location classifier just to discard users from the data set that belong to non-relevant locations in the study.

Once a classifier is trained with profiles that self-disclosed their attributes, we proceed to predict those attributes for the rest of the dataset, enabling us to have demographic attributes for all users. Note that XGBoost includes prediction confidence (where 1 is maximum confidence), and thus, we only keep users with a minimum confidence of 0.7 (determined by manual exploration). Technical details about this classification process can be found in Ref. [90].

3.3 TOPICS IN TRANSPORT MODELING

Some users may comment on their daily experiences, whether they are positive or negative, while other users participate in the political discussion of transport, giving their opinion on the subject. Therefore, this work will focus on classifying transport discussion into two main topics: *personal experience* and *general discussion*.

Given that the text from Twitter is unstructured, the task at hand requires a method to infer the latent structure that separates both aspects of content. We resort to topic modeling methods toward that aim. Particularly, we use Topic-Supervised Non-Negative Matrix Factorization (TS-NMF) [16], which has been used successfully in our previous work to infer the main mode of transportation of Twitter users [41], as well as mobile phone users [17].

Topic modeling techniques require a parameter with the number of topics, *i.e.*, the number of latent dimensions in the structure inferred by the model. In general, because these topics are latent, they do not always align with human interpretation. The TS-NMF method is semi-supervised, which means that each latent dimension can be guided toward a specific meaning. A set of terms can be pre-labeled with their association to one or more topics, which are two in our case. Then, the model guides the topic extraction process by aligning the obtained latent topics with the input information.

TS-NMF works by decomposing a known matrix D into the multiplication of two lower-rank matrices W and H of rank k. The supervision matrix L contains the available labels for user-documents (*i.e.*, rows of X). The W and H matrices are found by solving the following optimization problem:

$$\min_{W \geq 0, H \geq 0} ||D - (W \circ L)H||_F^2, \tag{1}$$

where $||\cdot||$ is the Frobenius norm, $\circ$ is the Haddamard product operator, k represents the number of topics, W represents the relevance of each topic in each document, and H represents the relevance of each term in each topic. The matrix L contains the subset of terms categorized into a topic, where each row is a term, and each column is a topic, and a cell value of 1 indicates association to the topic.

In our context, $k = 2$, and D is a term-user matrix, where each cell contains the weighted number of times the corresponding user published a tweet with the corresponding term. The weight is estimated using Term-Frequency–Inverse-Document-Frequency (TFIDF) [91], which measures how important a word is to a document in a collection or corpus. We apply TFIDF to D to give more weight to words that are less frequent in the data set, but that may be relevant for each discussion topic. The matrix L contains the subset of terms categorized into personal experience and general discussion. See Figure 7 for an overview of this approach.

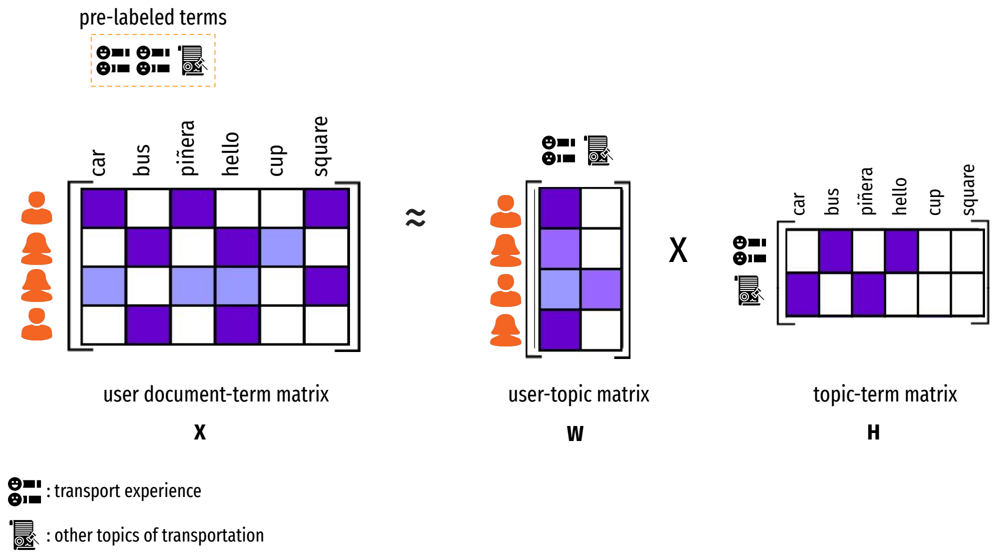

Figure 7: Depiction of the Topic-Supervised Non-Negative Matrix Factorization technique.

After applying the TS-NMF method, we obtain two matrices, one with the association of users with the topics of discussion, and one with the association of terms with the topics of discussion. We assign each user to a specific discussion group by estimating the maximum topic association. Users with a max topic score lower than 0.8 are discarded from the analysis due to being hard to classify (the threshold was determined via manual exploration).

Thus, at this stage of the methodology, for each user we know his/her/their gender and main topic of discussion.

3.4 TRANSPORT PERCEPTION METRIC

Once users have demographic attributes and the discussion is divided in two topics, the next step is to measure perception in each discussion. To do so, we follow a lexical approach, because language style can be an indicator of the subjective way in which people understand themselves and the different situations they take part in. Particularly, we use a well-known dictionary approach based on the psycho-linguistic dictionary named Linguistic Inquiry and Word Count (LIWC) [27], which is designed to capture emotional, cognitive, and structural components present in speech. LIWC defines several categories and words belonging to these categories. The frequency of word usage, then, is assumed to be a proxy of how these categories are related to who is performing the speech. As relevant categories for our study, we consider affective processes such as *anxiety*, *anger*, and *positive feelings*; relativity notions of *time* and *space*; and usage of certain words such as *cursing*, or

Table 3: LIWC Term Samples. The terms of the LIWC categories reflect cognitive processes such as emotions, and aspects related to transport.

Category	Sample Terms
Anger	*damn, assaulted, run over, aggressiveness, shit, threaten, motherfucker*
Anxiety	*riots, sorrows, scandalous, miserable, insecurity, worry, intimidation, dangerous*
Positive Feelings	*care, good, smile, want, applaud, happy, incentive, humor, wish, admire*
Time	*stopped, start, semester, minute, tip, june, late, season, we finish, start*
Space	*block, blocks, small, interior, region, zones, moving forward, base, through*
Sexual	*sexually, ass, sexual, naked, butt, darlings, fags, grabbed, rape*
Swear	*crest, assholes, faggot, asshole, fuck, shit, damn, motherfucker*

sexual (see Table 3 for examples of words from these categories). For this project, we updated the original dictionary to include Chilean slang, *e.g. weón* (dude, in *swear*), *agarrón* (grab, in *sexual*), and *taco* (traffic jam, in *space*).

We refer to our selected categories as *perception categories*, because they are proxies of how people perceive the transportation experience and the general discussion. We formalize this notion

The perception of users is represented in a Perception Matrix P using the LIWC dictionary, such that:

$$P(i,c) = \frac{\text{\# of times user i has used words from category c}}{\text{\# of words used by user i in any LIWC category}}, \tag{2}$$

where c is the perception category of interest, i is a user id. In this way, P contains emotional and contextual cues of how users feel regarding transportation, according to the frequency of words in perception categories. This matrix is divided into subgroups for the next step. For instance, two sub-matrices can be defined according to the most associated topic of users, P_T (subset of P with users associated with personal experience) and P_G (subset of P with users associated to general discussion). Four sub-matrices can be defined according to that definition, this time separating each group by gender.

Following previous work on perception in Twitter [92], we define the Transport Perception Metric TP of user i in sub-group S (*e.g.*, S is one of P_T or P_G) for category c as:

$$TP(i,c) = \frac{S(i,c) - \mu(S_c)}{\sigma(S_c)}, \tag{3}$$

where $\mu(S_c)$ is the mean of column vector S_c, and $sigma(S_c)$ is the standard deviation of column vector S_c. In summary, TP is a standardized measure of how each user in a subgroup expresses each category in relation to the rest of users in the same group.

Considering the previous definition of TP, we define the *gender gap in perception* as the difference in TP in a category c for a given discussion group S as:

$$GAP(S,c) = TP(S_f,c) - TP(S_m,c), \tag{4}$$

where S_f is the group of women users in S, and S_m is the group of men users in S. As result, GAP tells us if the tendency of using a specific perception category in a group is more skewed towards women (GAP > 0) or men (GAP < 0). Given that the GAP metric allows us to calculate the gender differences in each perception category, it allows us to answer our research question.

In summary, by following this methodology, it is possible to infer users' attributes, identify topics of discussion in transportation, estimate their perception in the discussion, and measure the gender differences of these perceptions. In the next chapter, we apply this methodology to tweets describing transportation in Santiago.

4 RESULTS

In this chapter, we present the results of a case study of transportation tweets in Spanish from Santiago, Chile. The main goal is to characterize the gender differences in transport perception as seen on Twitter.

4.1 DEMOGRAPHIC CLASSIFICATION

First, we report the results on demographic classification. To assess the prediction of gender, we evaluated the classifier with 5-fold cross-validation. The average precision is 0.712, a value that is below the state of the art of 0.967 [93]. To maintain the quality of the results, we only considered users that were classified with confidence. Note that our classifier relied on self-reported demographic found in users profiles, while the state of the art classifier has a more vast training set than ours, including the profile image of each user. We manually explored the dataset and found no relevant errors in gender prediction. Gender was determined by the usage of gendered-words in the biography, such as mom, engineer (man), and lawyer (woman); use of emojis, and the mention of words belonging to causes and sports (see Table 4). We classified 35% of users as women, and 65% as men. Comparing with the Chilean population, we found that 11 women and 20 men are present in our dataset for every 1,000 inhabitants. Two-thirds of users talking about transportation are men. The over-representation mirrors previous studies of demographic on the social platform [94], [95].

4.2 TOPICS IN TRANSPORTATION DISCUSSION

Next, we proceeded to infer the topic structure. As the first step, we built the labeling matrix L using a set of manually crafted seed terms. See example terms for each topic in Table 5.

Regarding personal experience, we assumed that people would mention specific terms such as station names, transport operators support accounts, highway names, ride-share apps like Uber or Cabify, and terms related to cycling and walking (see Quotes 4 and 5 from Table 2). For the general transportation discussion, we expected to identify users that discuss specific events, comment on publications from media outlets, send tweets directed to politicians, or comment about transportation policy, among other themes (see Quotes 6 and 7 from Table 2 for sample tweets).

Table 4: Gender classification metrics: precision and recall, and the top predictive features identified by the classifier.

	Precision	Recall	Top Predictors
Gender	0.712 ± 0.02	0.706 ± 0.02	profile:# of emojis, profile:mother (*madre*), profile:mom (*mamá*), profile:engineer (man, *ingeniero*), [words regarding political causes], profile:lawyer (man, *abogado*), [words about sports], profile:in love (woman, *enamorada*), profile:woman (*mujer*), profile:teacher (woman, *profesora*), profile:teacher (man, *profesor*), profile:chilean (man, *chileno*), profile:lawyer (woman, *abogada*), profile:crazy (woman, *loca*), profile:engineer (woman, *ingeniera*)

Table 5: Sample keywords used to guide the topic modeling of transportation discussion.

Topic	Sample keywords
Personal Experience	bus, subway (*metro*), #line5 (*#linea5*), 🚌, @metbus, bicycle, @bikesantiago, #ipedal, 🚴, highway (*autopista*), taxi, wazers (*community of Waze users*), #vespuciosur (a highway in Santiago), cabify (a ride-hailing application)
General Discussion	#census (*#censo*, population census in 2017), #chileelections, (*#eleccioneschile*, national elections), #chilevotes (*#chilevota*, national elections), @maratonsantiago (a sports event) #bienvenidos13 (a TV show), #lacooperativa (news media hashtag), @elmercurio_cl (news media), @adnradiochile (news media), @mop_chile (Ministry of Public Infrastructure), @mtt_chile (Ministry of Transportation), @paolatapiasalas (Ministry of Transportation), @marcoporchile (politician), @danieljadue (politician), @camaradiputados (Chamber of Deputies)

To assess the accuracy of the topic modeling method, we manually labeled a set of 389 users classified under each topic (see Table 6 for performance metrics). The precision in *personal experience* is 0.93 (with recall of 0.64), whereas the precision in *general discussion* is 0.40 (with recall of 0.83). The experience topic is correctly predicted, however, there are *personal experience* users that are being classified into *general discussion*. This is not bad *per se*, as our focus is the transport experience. To qualitatively explore how the topic extraction worked, we calculated a tendency value for each term in the dataset, defined as the difference between topic associations for each term. Figure 8 shows word clouds of terms with the highest tendency for each topic. Personal experience terms belong mostly to public transportation (subway –*metro*–, line –*línea*–, @metrodesantiago) and often also convey part of the interactions between service providers and users (see Quote 8 from Table 2). They often refer to the functioning and availability of the system, with words such as working (*funciona*), available (*disponible*), problems (*problemas*), service (*servicio*), and thank you (*gracias*). Meanwhile, terms related to general discussion consist mostly of mentions to politicians and service administrators (see Quote 9 from Table 2), as well as media outlets (@biobio) and politicians (@sebastianpinera, president of Chile).

Table 6: Manual evaluation of the topic model.

Topic	Precision	Recall	F1-Score
Personal Experience	0.93	0.64	0.76
General Discussion	0.40	0.83	0.54

(a) Personal Experience **(b)** General Discussion

Figure 8: Word clouds of the most associated terms to each topic of transportation.

The frequency of tweeting among topics shows resemblance to the general tweeting frequency, where transport discussion is higher during the weekdays to decrease over the weekend (see Figure 9). The relationship between both topics, in relative terms, is maintained during the whole week.

With respect to gender, a difference appears in the distribution per topic. In general, men represent a larger portion of both topics, and participate more in the general discussion, which means that men use Twitter as a platform to express their opinions on politics, event discussion, among others. We expected a majority of Twitter users to be male due to the relative demographics of the population in Chile, and therefore even topics more likely to be gendered towards women have an greater representation by male users" [96]. However, in proportion, women appear to talk more about their daily experiences when referring to transportation (see Figure 10).

4.3 TRANSPORT PERCEPTION

Following our methodology, the next step was estimating transport perception using LIWC categories in each discussion topic. As an exploratory analysis, we estimated the correlations between each category in the perception of users as defined in matrix P (see correlation matrices in Figure 11). The correlation between two different categories could help us understand the interactions between each of them. Time and space have a negative correlation in both personal experience and the general discussion of transport. This may be due to users only referring to one relative component in general, probably the one with the greatest personal importance. In that sense, time and positive feelings correlate negatively, which implies that users will not often

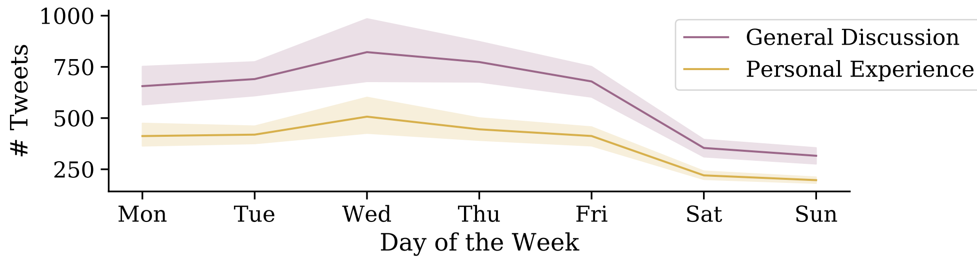

Figure 9: Tweet frequency per day of week, by topic of interest.

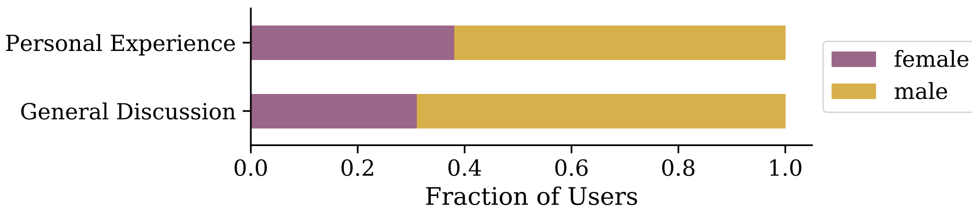

Figure 10: Distribution of users per topic of transportation and gender.

express themselves positively while referring to travel times (see Quote 10 from Table 2). On the other hand, the general transport discussion has its focus on the space-related component. This can be caused by users commenting on the design or planning of transport, commenting on vehicle congestion, or the little space inside the train or bus cars, among other things (see Quote 11 from Table 2).

Regarding emotional components, the use of words related to anxiety and anger are positively correlated in the personal experience topic, which is an expected result when people experience uncertainty in their travel. Positive feelings are negatively correlated with the space component for the personal experience discussion, perhaps related to the discontent users experience when they do not have enough space. Swear words are positively correlated with anger for both topics, an arguably expected result.

Then, we proceeded to estimate the differences with respect to gender in the overall discussion (see Figure 12). We found that women showed greater use of positive and anxious words, while men used more swear words. Other insights also emerged, with women using more words related to time, and words from the sexual category (See Quotes 12, 13, 14, and 15 from Table 2). Evidence shows that time plays an important role in the experience of women and their interaction with transportation. There is also greater use of words in the sexual category depending on the time. These results are consistent with previous results on the issue [97].

Finally, we estimated the differences in perception by gender for the two topics of discussion. Women expressed anxiety and positive feelings more frequently than men, in both dimensions of transport, skewing more to time. They also use sexual-related words, likely participating in the discussion about harassment. This indicates that while harassment is something they experience

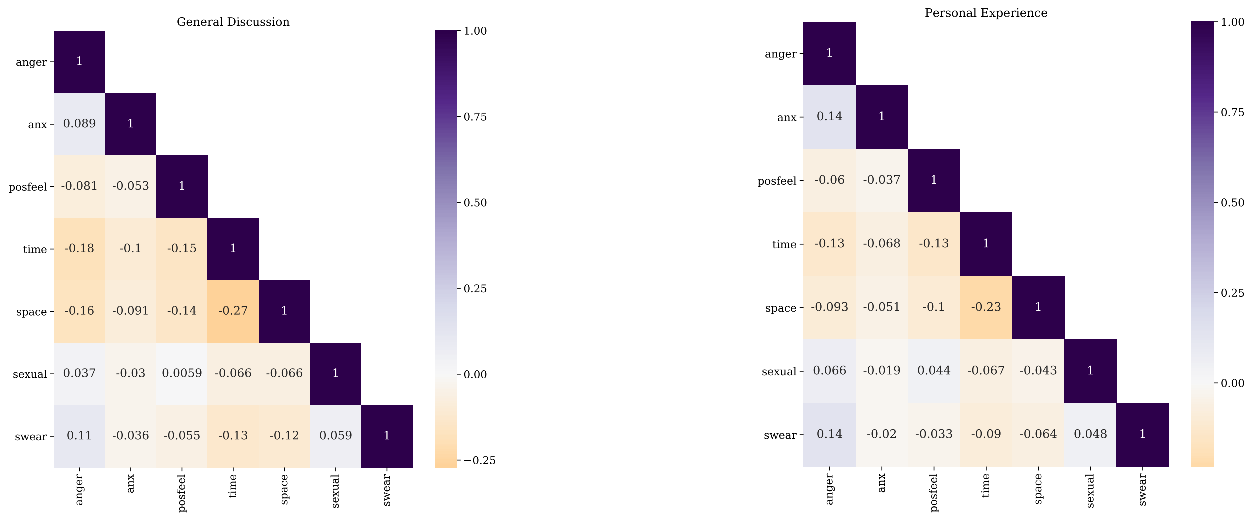

Figure 11: Matrix of significant correlations between Perception categories for the Personal Experience topic. All correlations are significant ($p < 0.05$, corrected using the Holm–Sidák method)

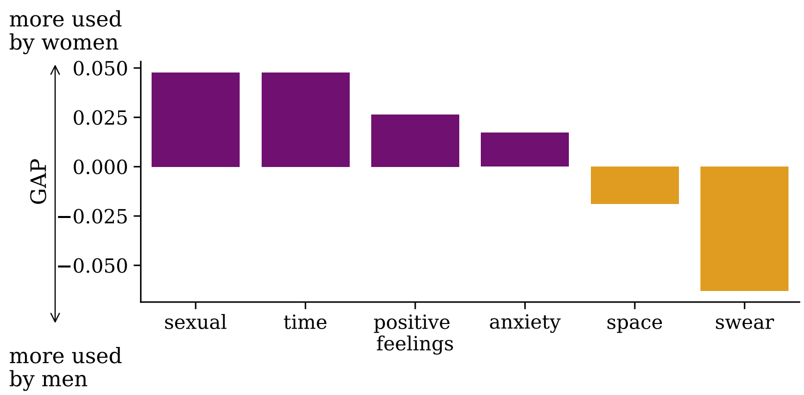

Figure 12: Gender Gap in Perception.

on a daily basis, it is seen as a systemic problem within transportation, which is also discussed from the approach of politics and planning. There are no differences in anger on the general discussion, but women express more anger in reference to their transportation experience, implying that there are aspects of transport where users do not feel satisfied. Also, women use more words related to time when talking about their personal experience, which indicates the relevance of this factor. We validated these results following a null model approach, where we estimated 1K tests shuffling the gender attribute of profiles. Figure 13 shows the comparison between the distribution of observed values (Effect Model, with distribution estimated using bootstrap sampling with replacement) and the null model.

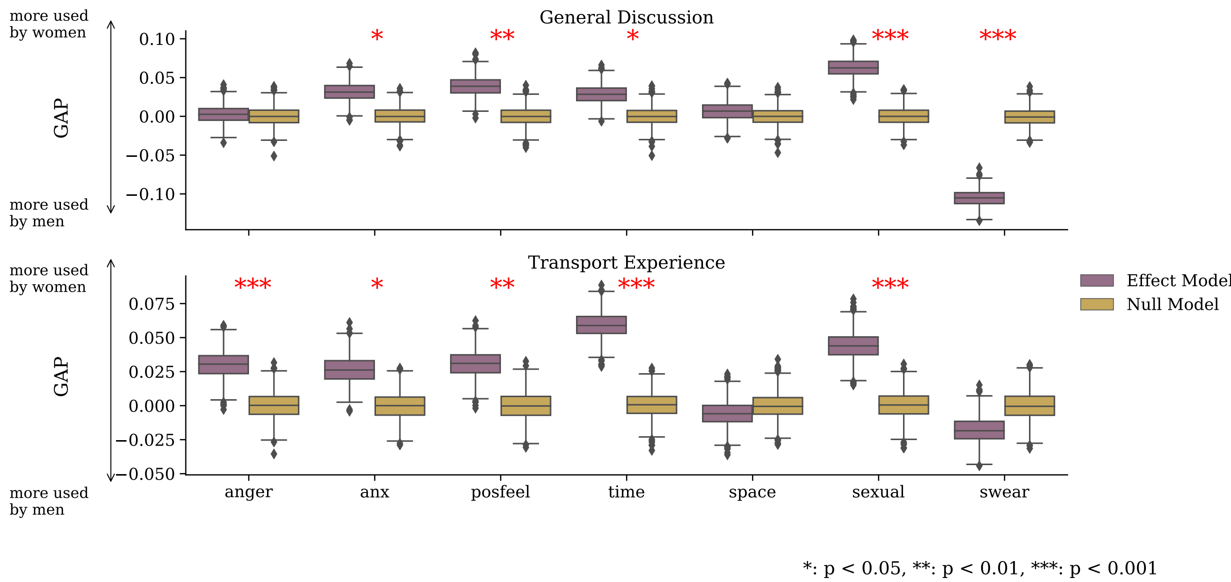

Figure 13: Boxplots of the GAP distributions per gender, topic, and perception category. Each plot contains the observed distribution (estimated with bootstrap sampling) and the null model (estimated with feature permutations).

4.4 COMPARISON WITH PREVIOUS WORKS

A previous study on gender differences in language use has shown that women often tend to express anxiety or positive feelings more frequently, with no differences in anger word usage [97]. However, for the personal experience discussion, women showed significantly more anger in their publications, proving their dissatisfaction with their journey experience. At the same time, women also referred more to time while talking about their personal experiences, which is consistent with previous studies on the perception of quality factors in which women place a high emphasis on punctuality in transportation systems [77]. The most significant gender difference is related to harassment to women, not only on transportation but also on the public space. Both issues are consistent with recent studies in Latin American cities [98].

Our results contradict previous assessments of sentiment toward public transport in Santiago de Chile, where it was found that most messages regarding transport experience were negative [85]. From a dataset point of view, both approaches are not directly comparable, as the referenced study focused on tweets about bus stops and bus routes, which may prompt adverse reactions when published on the go. From a methodological point of view, the study did not consider differences concerning gender, which may influence results considering the over-representation of men in Twitter [21], [96]. Addressing this disagreement could be done by defining more specific topics in future work.

5 DISCUSSION & CONCLUSIONS

The main aim of this dissertation was to design a methodology to measure the gender differences in transport perception as seen on social media platforms. We measured the gender differences across seven perception categories: affective processes (anxiety, anger, and positive feelings), relativity notions (such as time and space), and the use of words associated with cursing or sexual issues. In particular, we conducted a case study of gender differences in perception in micro-blogging messages from Santiago, Chile. Here we showed that both genders express differently. On the one hand, female users write about their emotions regarding travel (both positive and negative feelings). On the other hand, male users use more slang, making it difficult to interpret emotion. The most notable difference is related to female sexual-related harassment, not only on transportation, but also on the public space. This may indicate that women are more exposed to sexual harassment in transportation, a situation reported in other big cities in Latin America [14].

Overall, we believe that data obtained from social media can complement traditional information gathering methods, overcoming restrictions such as time and cost, while also allowing for previously unseen granularity of data. In this context, here we discuss the implications, limitations, and future work derived from this dissertation.

5.1 IMPLICATIONS

Social media analysis enables the identification of significant gender differences in the transport experience. Historically, these differences have been hidden due to the consideration of the "average user" in transportation. This identification of differences may be helpful for service providers and policymakers, not only in terms of management and planning, but also to improve accessibility of the public transportation with a gender perspective, in line with the Sustainable Development Goals by the UN, which aim to make cities more inclusive for all.

Service providers could track the most used terms to refer to transportation on their social networks, and study their occurrence based on events, time or characteristics of their users. Our method would allow them to highlight particularities that a survey or traditional method fails to show.

Another benefit of having gender indicators is that it makes it possible to monitor the situation of women in transport. It allows evidence-based decisions to be made by policymakers to design, formulate, monitor, and evaluate interventions effectively. We focused our analysis on Santiago, Chile, however, harassment in the public space and transportation is a common problem in Latin

America, thus, our work could be generalized and applied to other cities such as Quito, Buenos Aires, among others [98].

5.2 SCOPE AND LIMITATIONS

Critics may point out three aspects that scope the reach of this work: representativeness, geographical coverage, and methodological points of improvement.

First, we measured and compared transportation discussion with respect to gender, however, there are also other population groups with special needs, such as children and the elderly. Including them in this kind of study is not just a matter of adding categories to the demographic classifier, as these groups are also under-represented in social media in general. This aspect does not hinder the insights obtained: as long as it is made clear who is represented by the study, then insights are valuable and actionable.

Second, the lack of geographic information prevents us from comparing different sectors of the city, which may be crucial for transportation discussion. For instance, comparing the wealthier part of the city with deprived sectors may reveal different emotional patterns with respect to the transportation experience.

And third, our methodology can be deepened in several steps. For instance, users are assigned to their most associated topic, whereas a more advanced alternative would be to assign weights to topics. The consideration of the temporal aspect of discussions is also needed, as it would allow answering questions such as "what is the time of the day perceived as more safe/dangerous for women?" Still, our work has proven that it is possible to move in this direction. We leave these advancements to future work.

5.3 FUTURE WORK

In addition to address the limitations of this work, we devised two main lines of research: visualization and bias mitigation. Our current results provide insights regarding the transportation discussion, with identified implications for two types of stakeholders. However, it is not clear how to actually *transfer* those insights to these stakeholders in a meaningful way. Visualization has been a successful medium to communicate between data scientists and transportation experts, particularly in scenarios of solving transportation problems with non-traditional data sources and methods [99]. Finally, social media data is not considered a representative source of population information [21]. Several access gaps to technology, plus the different algorithmic and systemic biases in operation in these scenarios, imply that designing and implementing bias mitigation strategies are needed, especially if the insights derived are planned to be used in decision making and policy design.

5.4 FINAL WORDS

In this dissertation we quantified the gender differences in perception of transportation from a two-fold perspective: the personal experience of transport, and the general discussion around it. The usage of social media data, particularly from Twitter, which is not used traditionally for these issues, analyzed with methods from machine learning and psycho-linguistics, proved fruitful in finding topic- and gender-specific insights.

If these differences are taken into account by relevant stakeholders in city planning and management, policies could be implemented that help eliminates gender gaps in terms of access to work, education, opportunities, and above all, equal access to the opportunities a city has to offer, in line with the Sustainable Development Goals. Future work should address methodological aspects of our proposal, but also communication and transfer strategies to ensure its adoption in planning, complementing survey-based information, making visible critical problems that are hidden in these traditional instruments.

REFERENCES

[1] C. Cirillo, L. Eboli, and G. Mazzulla, "On the asymmetric user perception of transit service quality," *International Journal of Sustainable Transportation*, vol. 5, no. 4, pp. 216–232, 2011.

[2] H. Best and M. Lanzendorf, "Division of labour and gender differences in metropolitan car use: An empirical study in cologne, germany," *Journal of Transport Geography*, vol. 13, no. 2, pp. 109–121, 2005.

[3] S. Rosenbloom, "Trends in women's travel patterns," in *Women's Travel Issues Second National ConferenceDrachman Institute of the University of Arizona; Morgan State University; Federal Highway Administration*, 1998.

[4] N. McGuckin and E. Murakami, "Examining trip-chaining behavior: Comparison of travel by men and women," *Transportation research record*, vol. 1693, no. 1, pp. 79–85, 1999.

[5] J. Garrard, G. Rose, and S. K. Lo, "Promoting transportation cycling for women: The role of bicycle infrastructure," *Preventive medicine*, vol. 46, no. 1, pp. 55–59, 2008.

[6] J. L. Nasar, "Safe cities: Guidelines for planning, design, and management," *American Planning Association. Journal of the American Planning Association*, vol. 62, no. 1, p. 142, 1996.

[7] A. Loukaitou-Sideris, "Is it safe to walk?1 neighborhood safety and security considerations and their effects on walking," *Journal of Planning Literature*, vol. 20, no. 3, pp. 219–232, 2006. DOI: 10.1177/0885412205282770.

[8] A. Loukaitou-Sideris and C. Fink, "Addressing women's fear of victimization in transportation settings: A survey of us transit agencies," *Urban Affairs Review*, vol. 44, no. 4, pp. 554–587, 2009.

[9] A. Bastian and M. Börjesson, "The city as a driver of new mobility patterns, cycling and gender equality: Travel behaviour trends in stockholm 1985–2015," *Travel behaviour and society*, vol. 13, pp. 71–87, 2018.

[10] Y. Popuri, K. Proussaloglou, C. Ayvalik, F. Koppelman, and A. Lee, "Importance of traveler attitudes in the choice of public transportation to work: Findings from the regional transportation authority attitudinal survey," *Transportation*, vol. 38, no. 4, pp. 643–661, 2011.

[11] A. J. Richardson, E. S. Ampt, and A. H. Meyburg, *Survey methods for transport planning*. Eucalyptus Press Melbourne, 1995.

[12] L. Dell'Olio, A. Ibeas, and P. Cecin, "The quality of service desired by public transport users," *Transport Policy*, vol. 18, no. 1, pp. 217–227, 2011.

[13] L. Gauvin, M. Tizzoni, S. Piaggesi, A. Young, N. Adler, S. Verhulst, L. Ferres, and C. Cattuto, "Gender gaps in urban mobility," *ArXiv preprint arXiv:1906.09092*, 2019.

[14] G. Kash, "Always on the defensive: The effects of transit sexual assault on travel behavior and experience in colombia and bolivia," *Journal of Transport & Health*, vol. 13, pp. 234–246, 2019.

[15] F. Foundation, "Safe and sound: International research on women's personal safety on public transportation," FIA Foundation, Tech. Rep., 2016. [Online]. Available: https://www.fiafoundation.org/connect/publications/safe-and-sound.

[16] K. MacMillan and J. D. Wilson, "Topic supervised non-negative matrix factorization," *ArXiv preprint arXiv:1706.05084*, 2017.

[17] E. Graells-Garrido, D. Caro, and D. Parra, "Inferring modes of transportation using mobile phone data," *EPJ Data Science*, vol. 7, no. 1, p. 49, 2018.

[18] Huawei, "Global connectivity index," Huawei, Tech. Rep., 2020. [Online]. Available: https://www.huawei.com/minisite/gci/assets/files/gci_2019_whitepaper_en.pdf.

[19] E. Graells-Garrido, D. Caro, O. Miranda, R. Schifanella, and O. F. Peredo, "The www (and an h) of mobile application usage in the city: The what, where, when, and how," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1221–1229, ISBN: 9781450356404. DOI: 10.1145/3184558.3191561. [Online]. Available: https://doi.org/10.1145/3184558.3191561.

[20] R. T. Ilieva and T. McPhearson, "Social-media data for urban sustainability," *Nature Sustainability*, vol. 1, no. 10, pp. 553–565, 2018.

[21] R. Baeza-Yates, "Bias on the Web," *Communications of the ACM*, vol. 61, no. 6, pp. 54–61, 2018.

[22] E. Graells-Garrido, R. Baeza-Yates, and M. Lalmas, "How representative is an abortion debate on twitter?" In *Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 133–134.

[23] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[24] B. Pang, L. Lee, *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[25] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 347–354.

[26] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREc*, vol. 10, 2010, pp. 1320–1326.

[27] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[28] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared facebook data," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, ACM, 2014, pp. 626–638.

[29] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*, 2013.

[30] G. Harman and M. H. Dredze, "Measuring post traumatic stress disorder in twitter," *In ICWSM*, 2014.

[31] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Fourth international AAAI conference on weblogs and social media*, 2010.

[32] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[33] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[34] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, "Experimental explorations on short text topic mining between lda and nmf based schemes," *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.

[35] S. Wakamiya, R. Lee, Y. Kawai, and K. Sumiya, "Twitter-based urban area characterization by non-negative matrix factorization," in *Proceedings of the 2015 international conference on big data applications and services*, 2015, pp. 128–135.

[36] P. Lahoti, K. Garimella, and A. Gionis, "Joint non-negative matrix factorization for learning ideological leaning on twitter," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 351–359.

[37] W. Yang, L. Mu, and Y. Shen, "Effect of climate and seasonality on depressed mood among twitter users," *Applied geography*, vol. 63, pp. 184–191, 2015.

[38] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[39] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics, 2009, pp. 248–256.

[40] Y. Freire-Vidal and E. Graells-Garrido, "Characterization of local attitudes toward immigration using social media," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 783–790.

[41] P. Vasquez-Henriquez, E. Graells-Garrido, and D. Caro, "Characterizing transport perception using social media: Differences in mode and gender," in *Proceedings of the 10th ACM Conference on Web Science*, ACM, 2019, pp. 295–299.

[42] T. W. Andreassen, "(dis) satisfaction with public services: The case of public transportation," *Journal of Services Marketing*, 1995.

[43] M. L. Tischer and R. V. Phillips, "The relationship between transportation perceptions and behavior over time," *Transportation*, vol. 8, no. 1, pp. 21–36, 1979.

[44] S. G. Stradling, J. Anable, and M. Carreno, "Performance, importance and user disgruntlement: A six-step method for measuring satisfaction with travel modes," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 1, pp. 98–106, 2007.

[45] M. Fellesson and M. Friman, "Perceived satisfaction with public transport service in nine european cities," in *Journal of the Transportation Research Forum*, vol. 47, 2012.

[46] G. Beirão and J. S. Cabral, "Understanding attitudes towards public transport and private car: A qualitative study," *Transport policy*, vol. 14, no. 6, pp. 478–489, 2007.

[47] M. Diana, "Measuring the satisfaction of multimodal travelers for local transit services in different urban contexts," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 1, pp. 1–11, 2012.

[48] H. Iseki and B. D. Taylor, "Style versus service? an analysis of user perceptions of transit stops and stations," *Journal of Public Transportation*, vol. 13, no. 3, p. 2, 2010.

[49] M. Bordagaray, A. Ibeas, and L. dell'Olio, "Modeling user perception of public bicycle services," *Procedia-Social and Behavioral Sciences*, vol. 54, pp. 1308–1316, 2012.

[50] D. T. Ory and P. L. Mokhtarian, "When is getting there half the fun? modeling the liking for travel," *Transportation Research Part A: Policy and Practice*, vol. 39, no. 2-3, pp. 97–123, 2005.

[51] J. Allen, J. C. Muñoz, and J. de Dios Ortúzar, "Modelling service-specific and global transit satisfaction under travel and user heterogeneity," *Transportation Research Part A: Policy and Practice*, vol. 113, pp. 509–528, 2018.

[52] Y. Tyrinopoulos and C. Antoniou, "Public transit user satisfaction: Variability and policy implications," *Transport Policy*, vol. 15, no. 4, pp. 260–272, 2008.

[53] J. Del Castillo and F. Benitez, "Determining a public transport satisfaction index from user surveys," *Transportmetrica A: Transport Science*, vol. 9, no. 8, pp. 713–741, 2013.

[54] L. Eboli and G. Mazzulla, "A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view," *Transport Policy*, vol. 18, no. 1, pp. 172–181, 2011.

[55] E. St-Louis, K. Manaugh, D. van Lierop, and A. El-Geneidy, "The happy commuter: A comparison of commuter satisfaction across modes," *Transportation research part F: Traffic psychology and behaviour*, vol. 26, pp. 160–170, 2014.

[56] C. Pronello and C. Camusso, "Travellers' profiles definition using statistical multivariate analysis of attitudinal variables," *Journal of Transport Geography*, vol. 19, no. 6, pp. 1294–1308, 2011.

[57] J. Anable, "'complacent car addicts' or 'aspiring environmentalists'? identifying travel behaviour segments using attitude theory," *Transport policy*, vol. 12, no. 1, pp. 65–78, 2005.

[58] M. Diana and P. L. Mokhtarian, "Grouping travelers on the basis of their different car and transit levels of use," *Transportation*, vol. 36, no. 4, pp. 455–467, 2009.

[59] E. Molin, P. Mokhtarian, and M. Kroesen, "Multimodal travel groups and attitudes: A latent class cluster analysis of dutch travelers," *Transportation Research Part A: Policy and Practice*, vol. 83, pp. 14–29, 2016.

[60] D. Ettema, T. Gärling, L. Eriksson, M. Friman, L. E. Olsson, and S. Fujii, "Satisfaction with travel and subjective well-being: Development and test of a measurement tool," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 3, pp. 167–175, 2011.

[61] B. Gatersleben and D. Uzzell, "Affective appraisals of the daily commute: Comparing perceptions of drivers, cyclists, walkers, and users of public transport," *Environment and behavior*, vol. 39, no. 3, pp. 416–431, 2007.

[62] L. Eriksson, M. Friman, and T. Gärling, "Perceived attributes of bus and car mediating satisfaction with the work commute," *Transportation Research Part A: Policy and Practice*, vol. 47, pp. 87–96, 2013.

[63] M. Sweet and P. Kanaroglou, “Gender differences: The role of travel and time use in subjective well-being,” *Transportation research part F: Traffic psychology and behaviour*, vol. 40, pp. 23–34, 2016.

[64] L. Mejıa-Dorantes and P. S. Villagrán, “A review on the influence of barriers on gender equality to access the city: A synthesis approach of mexico city and its metropolitan area,” *Cities*, vol. 96, p. 102 439, 2020.

[65] P. Gordon, A. Kumar, and H. W. Richardson, “Gender differences in metropolitan travel behaviour,” *Regional Studies*, vol. 23, no. 6, pp. 499–510, 1989.

[66] S. Hanson and I. Johnston, “Gender differences in work-trip length: Explanations and implications,” *Urban geography*, vol. 6, no. 3, pp. 193–219, 1985.

[67] R. Law, “Beyond ‘women and transport’: Towards new geographies of gender and daily mobility,” *Progress in human geography*, vol. 23, no. 4, pp. 567–588, 1999.

[68] O. Blumen, “Gender differences in the journey to work,” *Urban geography*, vol. 15, no. 3, pp. 223–245, 1994.

[69] S. Rosenbloom, *The need for study of women’s travel issues*, 1978.

[70] M. Mauch and B. D. Taylor, “Gender, race, and travel behavior: Analysis of household-serving travel and commuting in san francisco bay area,” *Transportation Research Record*, vol. 1607, no. 1, pp. 147–153, 1997.

[71] G. Guiliano, “Public transportation and the travel needs of women,” *Traffic Quarterly*, vol. 33, no. 4, 1979.

[72] R. Macmillan, A. Nierobisz, and S. Welsh, “Experiencing the streets: Harassment and perceptions of safety among women,” *Journal of Research in Crime and Delinquency*, vol. 37, no. 3, pp. 306–322, 2000. DOI: 10.1177/0022427800037003003.

[73] G. Lynch and S. Atkins, “The influence of personal security fears on women’s travel patterns,” *Transportation*, vol. 15, no. 3, pp. 257–277, 1988.

[74] J. Garrard, “Healthy revolutions: Promoting cycling among women,” *Health Promotion Journal of Australia Health Promotion Journal of Australia*, vol. 1414, pp. 213–5, Jan. 2003.

[75] N. Yavuz and E. W. Welch, “Addressing fear of crime in public space: Gender differences in reaction to safety measures in train transit,” *Urban Studies*, vol. 47, no. 12, pp. 2491–2515, 2010. DOI: 10.1177/0042098009359033.

[76] H.-P. Hsu, “How does fear of sexual harassment on transit affect women’s use of transit?” In *Proc., Women’s Issues in Transportation: Summary of the 4th International Conference*, vol. 2, 2010, pp. 85–94.

[77] M. Rojo, H. Gonzalo-Orden, L. Dell'Olio, and A. Ibeas, "Modelling gender perception of quality in interurban bus services," in *Proceedings of the Institution of Civil Engineers-Transport*, Thomas Telford Ltd, vol. 164, 2011, pp. 43–53.

[78] Y. Fan, A. Guthrie, and D. Levinson, "Waiting time perceptions at transit stops and stations: Effects of basic amenities, gender, and security," *Transportation Research Part A: Policy and Practice*, vol. 88, pp. 251–264, 2016.

[79] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*, Springer, 2011, pp. 338–349.

[80] Z. Zhang, Q. He, and S. Zhu, "Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 396–414, 2017.

[81] E. Chaniotakis, C. Antoniou, G. Aifadopoulou, and L. Dimitriou, "Inferring activities from social media data," *Transportation research record*, vol. 2666, no. 1, pp. 29–37, 2017.

[82] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.

[83] C. Collins, S. Hasan, and S. V. Ukkusuri, "A novel transit rider satisfaction metric: Rider sentiments measured from online social media data," *Journal of Public Transportation*, vol. 16, no. 2, p. 2, 2013.

[84] M. Congosto, D. Fuentes-Lorenzo, and L. Sánchez, "Microbloggers as sensors for public transport breakdowns," *IEEE Internet Computing*, vol. 19, no. 6, pp. 18–25, 2015.

[85] J. T. Méndez, H. Lobel, D. Parra, and J. C. Herrera, "Using twitter to infer user satisfaction with public transport: The case of santiago, chile," *IEEE Access*, vol. 7, pp. 60 255–60 263, 2019.

[86] D. Li, Y. Zhang, and C. Li, "Mining public opinion on transportation systems based on social media data," *Sustainability*, vol. 11, no. 15, p. 4016, 2019.

[87] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: The dynamics of the location field in user profiles," in *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, 2011, pp. 237–246.

[88] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 2016, pp. 785–794.

[89] W. Liu and D. Ruths, “What’s in a name? using first names as features for gender inference in twitter.,” in *AAAI spring symposium: Analyzing microtext*, vol. 13, 2013, pp. 10–16.

[90] E. Graells-Garrido, R. Baeza-Yates, and M. Lalmas, “Every colour you are: Stance prediction and turnaround in controversial issues,” in *Proceedings of the 12th ACM Conference on Web Science*, ACM, 2020.

[91] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[92] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, “Tracking gross community happiness from tweets,” in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, ACM, 2012, pp. 965–968.

[93] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens, “Demographic inference and representative population estimates from multilingual social media data,” in *The World Wide Web Conference*, 2019, pp. 2056–2067.

[94] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, “Understanding the demographics of Twitter users,” in *Fifth international AAAI conference on weblogs and social media*, 2011.

[95] P. Longley, M. Adnan, and G. Lansley, “The geotemporal demographics of twitter usage,” *Environment and Planning A: Economy and Space*, vol. 47, no. 2, pp. 465–484, 2015. DOI: 10.1068/a130122p.

[96] E. Graells-Garrido, R. Baeza-Yates, and M. Lalmas, “Representativeness of abortion legislation debate on twitter: A case study in argentina and chile,” in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 765–774.

[97] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, “Gender differences in language use: An analysis of 14,000 text samples,” *Discourse Processes*, vol. 45, no. 3, pp. 211–236, 2008.

[98] H. Allen, *Ella se mueve segura. un estudio sobre la seguridad personal de las mujeres y el transporte publico en tres ciudades de america latina*. Caracas: CAF y FIA Foundation, 2019. [Online]. Available: Caracas:%20CAF%20y%20FIA%20Foundation.

[99] E. Graells-Garrido and V. Peña-Araya, “Toward an interdisciplinary methodology to solve new (old) transportation problems,” in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 504–509.

Appendices

A | CONFERENCE PAPER

The following pages include the paper with the following reference: Vasquez-Henriquez, P., Graells-Garrido, E., & Caro, D. (2019, June). Characterizing Transport Perception using Social Media: Differences in Mode and Gender. In *Proceedings of the 11th ACM Conference on Web Science* (pp. 295-299).

This thesis is an extended version of the paper.

Characterizing Transport Perception using Social Media: Differences in Mode and Gender

Paula Vasquez-Henriquez
Universidad del Desarrollo
Santiago, Chile
pvasquezh@udd.cl

Eduardo Graells-Garrido
Universidad del Desarrollo
Santiago, Chile
egraells@udd.cl

Diego Caro
Universidad del Desarrollo
Santiago, Chile
dcaro@udd.cl

ABSTRACT

Transport planners face the growing need to understand the behavior of their users, who base their mobility decisions on several factors, including travel time, quality of service, and security. However, transportation is usually designed with an average user in mind, without considering the needs of important groups, such as women. In this context, we analyzed 300K tweets about transportation in Santiago, Chile. We classified users into modes of transportation, and then we estimated the associations between mode of transportation, gender, and the categories of a psycho-linguistic lexicon. Our results include that women express more anger and sadness than expected, and are worried about sexual harassment. Conversely, men focus more on the spatial aspects of transportation, leisure, and work. Thus, our work provides evidence on which aspects of transportation are relevant in the daily experience, enabling the measurement of the travel experience using social media.

CCS CONCEPTS

• **Information systems** → **Web mining**; • **Applied computing** → Transportation;

KEYWORDS

sentiment analysis; twitter; gender differences; transportation

ACM Reference Format:
Paula Vasquez-Henriquez, Eduardo Graells-Garrido, and Diego Caro. 2019. Characterizing Transport Perception using Social Media: Differences in Mode and Gender. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3292522.3326036

1 INTRODUCTION

Transportation plays a key role in people's development in society, greatly affecting the quality of life [7, 32] of its users, who base their mobility decisions on factors such as cost, comfort, accessibility, punctuality, quality of service, and security [10]. For this reason, transport policy-makers face the growing need to understand their needs and perceptions to better plan and manage transportation networks and public policy. However, there is a gap between the perceptions of transportation administrators and those from users, since information is collected with an "average" user in mind, usually with little consideration of the needs and opinions of other important user groups, such as women [30].

In this paper, we characterized the perception of transportation through Twitter, taking into account differences in mode of transportation and gender. We hypothesized that, if different modes of transportation have their own specific issues, or if women experience any mode of transportation in a different way than men, then this should be reflected in a difference in the linguistic components of the texts they publish. We measured these linguistic components in a two-step process. First, we classified users and their transportation-related tweets into mode(s) of transportation using semi-supervised topic modeling [9, 20]. Then, we measured perception using a psycho-linguistic lexicon and gross-community perception metrics [25].

We applied our proposed method to 300K tweets published in Santiago, Chile. We found that public transport users use Twitter to interact with service providers, while reporting a higher association with sexual and swear words. Motorized transport discussion focuses on taxis and ride-hailing apps, and the state of the driving system, where users show to be sensitive to space usage. Non-motorized transportation discussion is centered around the leisure aspect of this mode [21], where users report more optimism. In terms of gender differences, we confirmed the differences in perception and focus. Women are more associated with sociability, sexual harassment, and positive words, pointing to an ambivalence between concerns and positive experiences. Conversely, men are more associated with the spatial aspect of transportation, work discussion, and swear words.

Our work contributes a methodology to infer mode of transportation usage from social media content, and a case study of measured differences between modes of transportation, with a gender perspective, using Twitter data from a big city. Our results provide evidence on which aspects of transportation are relevant in the travel experience, as measured from social media. Our metrics could be put into operation to allow transportation planners to consider a wider range of needs and dynamics into their work, complementing traditional data sources with fine-grained data.

2 RELATED WORK

Data used to plan and manage transportation is traditionally collected through surveys [1, 8]. Women's travel patterns differ greatly from those of men due to the role they play in society, combining the tasks of workers, home-care givers and those responsible for children and the elderly [2, 19]. However, these differences often do not appear on survey data, where budget constraints of traditional

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326036

transportation questionnaires may not allow to incorporate gender-specific questions. As result, most of the planning is targeted at average users, with a lack of a gender perspective.

The role of social media in transport analysis has rapidly grown over the last years, allowing to obtain information regarding trips and activities, while highlighting the benefits of using this type of data sources [26]. Social platforms such as Twitter, and other specialized platforms like Waze, allow users to organize in communities, contributing information about transportation, which has been used to infer patterns and dynamics of urban behavior [14, 28]. To the extent of our knowledge, there is not much literature on alternative ways to approach transport perception, except for a descriptive analysis of public transport perception from tweets in the city of Chicago [5], and the monitoring of malfunctions in public transport in Madrid [6].

Our work can be seen as a deeper analysis on the travel experience than those from previous work [5, 6], due to our focus on all modes of transportation and to our structured lexical analysis.

3 METHODOLOGY

The main goal of this study is to characterize the perception of transportation as seen on Twitter, quantifying differences per mode of transport and gender. In this section, we explain the three steps of the pipeline we used to achieve this purpose: user representation and inference of gender, inference of mode of transportation, and measurement of perception.

User Features. A Twitter user profile contains the following attributes used in our study: *id*, *name*, *description*, and *tweets*. Tweets are textual micro-posts of 280 characters at most that may contain mentions to other users, hashtags to indicate themes within the post, URLs, emoji, *etc.* To these features, we also added *gender*. We focused on a binary gender separation: {male, female}. Without losing generality, we inferred gender based on self-reported information [12, 17]. The inference was based on two heuristics. The first one matched the first name of each user with a database of known names, built from census data and from manually crafted lists. The second one matched expressions in the self-reported description (*e.g.*, "Mother, Sister, Daughter.....", and so on). Then, we propagated gender labels using *SGD Classifier* implemented in *scikit-learn* [23]. We did so by predicting gender based on the description content.

Mode of Transportation Inference. To measure perception per mode of transportation, we needed a way to classify users and tweets into each mode. We used the following high-level categorization of mode(s) of transportation: 1) Public Transportation (*e.g.*, bus, subway); 2) Motorized Transportation (*e.g.*, cars, ride-apps); and 3) Non-Motorized Transportation (*e.g.*, bicycles). These modes can be characterized by vocabulary usage, for instance, public transportation operators have support accounts, and use specific terms like station names. Thus, it is possible to train a model that infers the relationships between users and modes, as well as terms and modes based on the co-occurrences of words. Based on a model to infer modes(s) of transportation from mobile phone data [9], we used a semi-supervised method named Topic-Supervised Non-Negative Matrix Factorization (TS-NMF) [20], which associates users and tweets to latent features interpreted as modes of transportation.

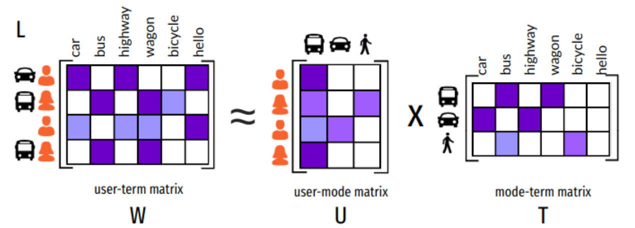

Figure 1: Topic-Supervised Non-Negative Matrix Factorization technique. The document-term matrix W holds the vocabulary used by each user in all their tweets, where each row represents an user, and columns represent the number of times a term is used. As W is positive definite, it can be decomposed in two matrices U and T of lower dimensions. These two matrices represent the latent features associations between terms and modes of transportation.

The TS-NMF method receives as input a document-term matrix W and a supervision matrix L. The matrix W was built from the concatenation of user timelines (*i.e.*, all tweets by the same user), which we treated as documents. A document d_u is defined as a vector: $d_u = [w_1, w_2, ..., w_{|V|}]$, where w_i represents the normalized frequency of term i posted by user u, and $|V|$ is the size of the vocabulary. Terms include words and n-grams (up to three), hashtags, mentions, URL hostnames, and emojis. The matrix L contains a subset of pre-labeled users with modes of transportation. For each mode, we had a list of seed words and phrases that represent it (*e.g.*, *platform* is associated with public transport, *Uber* is associated with motorized transport). Users were pre-labeled based on a calculated score defined as the sum of normalized frequencies of these seed words for each mode. Those with a score higher than a threshold were labeled with the corresponding mode (in our experiments, we defined 0.25 as a compromise between confidence and number of users labeled).

The TS-NMF method decomposes W into the product of two matrices of lower dimensions. This factorization allowed us to arrange the vocabulary into clusters (or latent components) according to the mode of transportation described by users. To do so, we decomposed the matrix into $W \approx U \times T$, where U was a $|u| \times k$ matrix that encodes k latent features of users (i.e their mode of transportation), and T is a $k \times |V|$ matrix, that encodes k latent features over the vocabulary (c.f. Fig. 1). The TS-NMF factorization takes into account the pre-labeled users in L to promote a meaningful semantic structure in the decomposition of the k latent features. For a deep review in topic-supervised factorization see [20].

In this way, the matrix U characterizes users, and the matrix T characterizes the vocabulary, allowing us to classify user tweets into modes of transportation according to their vocabulary usage.

Gross Transport Perception. We quantified transportation perception through the usage of the psycho-linguistic lexicon Linguistic Inquiry and Word Count (LIWC) [24]. This lexicon has been used to characterize perception and emotions on Twitter [11]. Particularly, it defines three high-level categories that we deemed relevant for our study: *Emotionality*, *Relativity*, and *Personal Concerns*. Emotionality includes both negative emotions such as *anger, sadness,* *fear and anxiety*, and positive emotions such as *optimism* and *positive feelings*. Relativity includes notions of *time, motion*, and *space*. Personal Concerns include themes such as *job, leisure, social, swear words*, and *sexual*-related words.

However, just counting words of each LIWC category is not enough, as social media content is subject to factors that may bias analysis. For instance, it can be expected to have more tweets about public transport than motorized modes given that transit riders may use their phones while moving. Previous work on Gross Perception [25] tackles this issue by analyzing the standardized usage of LIWC categories. We built upon these metrics by defining the Gross Transportation Perception GTP, which measured the relative use of words per each mode and LIWC category as:

$$\mathrm{GTP}_x^{p_{ij}} = \frac{x_{ij}^p - \mu_i^p}{\sigma_i^p},$$

where p is a LIWC category, i is a mode of transportation, j is a day, and x is the normalized frequency of words belonging to the category p in said day for a given group of tweets, μ_i^p and σ_i^p are the mean and sd of the fraction of words belonging to the category p of the mode of transport i. It may be a single user, a group of users, all tweets following specific criteria, *etc.* We carried out this analysis at the transport mode level, in periods of three hours, and at the gender level, at daily periods.

Having into account the previous definition, we defined the *gender gap in perception* as $GAP_i^p = GTP_{f,i}^p - GTP_{m,i}^p$, where p is a LIWC category, i is a mode of transportation, f is the set of tweets by women, and m is the set of tweets by men in a given period. As result, *GAP* tells us if the tendency of using a certain category is associated with females (GAP > 0) or males (GAP < 0).

4 CASE STUDY: SANTIAGO, CHILE

Santiago, Chile, is a densely populated city with almost 7 million people. We collected tweets related to transportation from March to November of 2017, leaving out the summer period in which mobility patterns change. We filtered out tweets that contain URLs of media outlets and unrelated topics, and users with a reported location different from the city of Santiago. We queried the Twitter Streaming API with words related to transportation, such as operator names, station names, and transport-application names. We collected a total of 303,800 tweets from 56,624 users living in the Santiago metropolitan area. We were able to identify 19,012 users as women, and 29,166 as men, which together represent over 85% of the sample. Note that users without inferred gender were later considered to measure differences in modes of transportation.

Figure 2 shows the aggregated hourly frequency distribution of tweets. One can see that the frequency resembles the morning and afternoon peak hours in transportation [33]. The most common terms are the institutional accounts of public transport services (@transantiago and @metrodesantiago). Such frequency hints that a big part of the discussion is held by public transport.

Mode of Transportation Inference. To associate modes of transportation to users and terms we defined seed keywords for each mode, including account names, hashtags, plain words, and emojis. With this schema, we pre-labeled 16,375 users to public transport, 2,449 users to motorized transport, and 2,447 users to non-motorized transport. The TS-NMF model propagated these labels to the rest of the data set (*c.f.* Table 1).

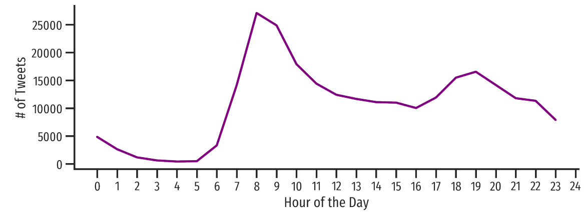

Figure 2: Tweet Frequency per hour of day. The transport-related tweeting frequency resembles the morning and the afternoon peak hours in the Santiago transportation system.

Mode	Seed Word	# Users	# Tweets
Motorized	*driving, highway, @uber_chile*	2,449	41,152
Non-Motorized	*bicycle, walking, @mobikecl*	2,447	29,051
Public	*@transantiago*, 🚌, *@metrodesantiago*	16,375	233,597

Table 1: Seed words used to characterize transportation mode. Some users were pre-labeled based according to their usage of the seed words of each mode of transportation.

Figure 3 shows one wordcloud per mode of transportation, with their corresponding associated words according to the TS-NMF model. The motorized transport discussion focused on terms related to taxis and ride-hailing apps (*@uber_chile, cabify, taxis, taxi drivers, #uber*), possibly due to conflicts about the legality of these services. Other related terms include the state of the driving system (*highway, driving, accident*), service providers (*@uoct_rm, @autopcentral*), and trip information via Waze, a community-driven GPS navigation software where users (who refer themselves as *wazers*) share travel times and route details. Non-motorized terms relate to riding the bicycle (*bicycle, cyclists, @mfc_stgo*) and their users show more relationship with accounts from municipalities (*@muni_stgo*) andauthorities (*@alessandrifelip*, the mayor of Santiago; or *@orrego*, the former city's intendant). This may be due to two things: people mentioning them to communicate an opinion or a retweet of information published by them. Terms related to public transportation are in respect to the Transantiago system (which integrates the buses and subway), with users mentioning station names (*e.g., Baquedano*, a downtown connection hub), and using hashtags related to the line services and service status (Lines 1 and 6, *#l1, #l6*). Also, they often interact with the Ministry of Transport and Telecommunications (*@mtt_chile*), or bus providers (*@alsaciaexpress*). They mention fewer accounts in comparison to other modes, but use more terms to describe the state and characteristics of the system (*day, minutes, wait, bus stop, wagons*).

Gross Transport Perception per Mode. Figure 4 shows the gross-perception per mode of transportation and LIWC categories. The emotionality associations show that *anger* is a less associated emotion with non-motorized transportation, which is consistent with studies that show that walking and cycling are more related to well-being and higher satisfaction [29, 31]. *Anxiety* is present in public transportation, which could indicate that there is a situation that triggers feelings of fear in users. *Optimism* signals a difference between the non-motorized and motorized transportation, where

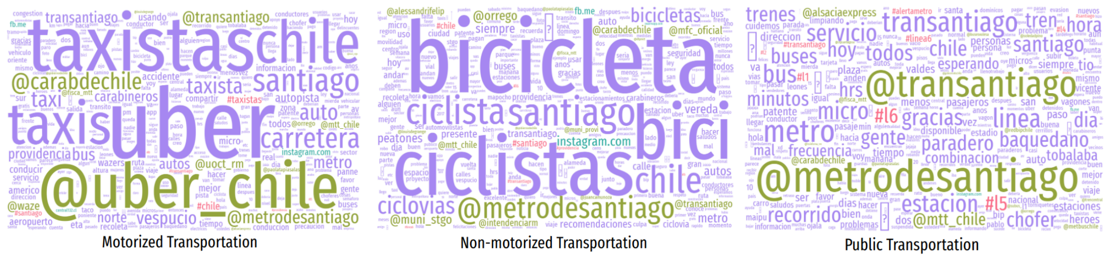

Figure 3: Wordclouds of the most associated words per mode of transportation. Motorized transportation discussion focuses on terms related to taxis, ride-hailing apps and the state of the driving system, while non-motorized transportation talk mainly about bicycles and public transportation focuses on interactions with service providers.

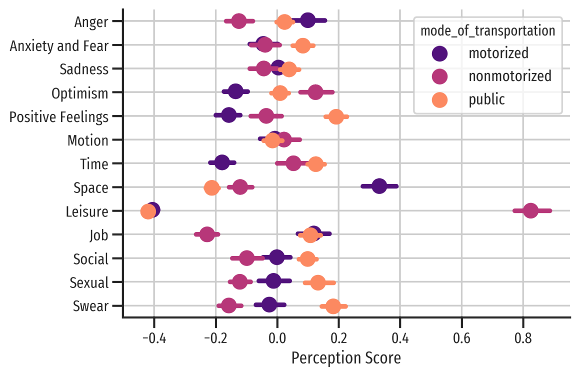

Figure 4: Gross Perception per Mode of Transportation (GTP), a measure of the relative usage of LIWC categories between modes of transportation. For instance, leisure is more associated with non-motorized transportation, in comparison to the other modes.

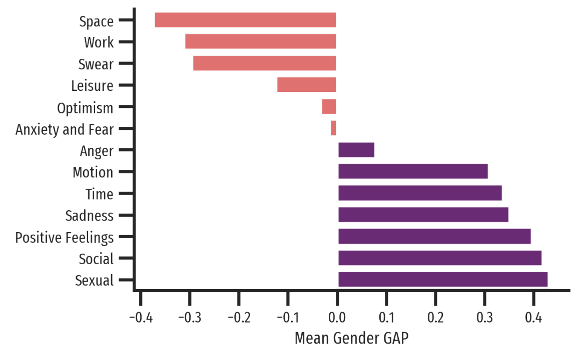

Figure 5: Mean Gender Gaps in Gross-Perception, calculated as the difference between GTPs. LIWC Categories with positive (negative) values present higher association with women (men). For instance, women report higher association to sexual words.

the latter seem to experience fewer feelings of optimism during their travels, which could be related to the stress drivers experience, as studies have shown [15]. *Positive feelings* are more associated to public transportation, contrary to the intuition that Twitter is used only for complains [13].

In terms of relativity, *time* is more associated with non-motorized and public transportation. For example, public transport users may talk about how long they have been waiting for a bus to arrive; non-motorized users may comment on their trip times. Motorized transport users pay more attention to *space*, due to issues such as the state of roads, traffic jams, and accidents. It is important to study these factors since users base their transportation decisions on them [10]. This applies even to single trips, due to the dynamic conditions of transportation, such as crowding at peak hours.

Regarding personal concerns, *leisure* is more associated with non-motorized transport. This may be related to the fact that non-motorized transportation has a strong use for entertainment purposes such as biking or taking a walk for pleasure [21]. In a similar way, public transport association to *work* emerges in the higher use of words from this feature, due to its high use for commuting. In Santiago, commuting represents more than a third of the total trips per day [33]. Public transport users seem to use more *sexual* and *swear* words. In addition to the various studies that have shown that public transport users present negative feelings during the trip [3]; factors such as crowding, delays, and accessibility increase stress, and that this perception differs according to factors such as gender [4]. It is in this mode of transport where people report high rates of sexual harassment [27].

Gender Differences. Figure 5 shows the mean gender gaps in gross-perception of LIWC categories, after estimating GTP for each gender. Men are more associated to the *space* category, which we theorize is because they tend to be the main users of motorized transport [33], where these terms are frequently used. Men are also associated with words related to *work* and *leisure*, consistent with results that show that they talk more about external events, objects, and processes, using more *swear* words [22].

On the other hand, women are more associated with words that talk about the context of transport (*motion, time*). As discussed, this may be due to factors such as time influencing their transportation decisions. It is notable that women also report a high association with *positive* words while also being highly associated with feelings such as *sadness* or *anger*, resulting in an ambivalence.We theorize that this may be caused by the use of sarcasm, or other reasons that should be explored in greater depth. Negative feelings could be related to their higher use of words from the *sexual* sphere, in accordance with studies that point to the violence they suffer in transportation [18], especially in the public type [27]. We explored the words that co-occur with those belonging to the sexual category, finding words such as *harassment, rape, street, man, sexual*, and mentions of public transport (*metro, transantiago, @transantiago, micro*), which is where women report suffering more harassment.

5 DISCUSSION AND CONCLUSIONS

Transportation is fundamental for the development in society, and the way it is perceived has a great impact on the quality of life. Hence, it is important to characterize the perception of people with respect to this activity. In this paper, we have established a method to quantify and compare the perception of modes of transportation, including gender as a factor of analysis. This method captures the subjective travel experience in an inexpensive and dynamic way. Entities such as service providers and transportation system administrators can benefit from the knowledge gained through social network data.

The creation of gender-aware subjective experience metrics could help to identify relevant issues regarding the travel experience that are ignored when transport is designed with the "average user" in mind. For instance, even though sexual harassment is a problem for women, the last travel survey in Santiago only referred to safety, potentially including harassment, but also confounding it with other unsafe situations [33]. This is of special importance in countries with high levels of gender inequality, where women are more often victims of sexual harassment in the public space [18]. Transport riding quality and satisfaction are often measured in terms of the needs of the business [5], which may be oblivious to social problems.

It could be argued that the use of social media data may be biased in terms of representativeness. Although the proportion of men and women may not be representative of the population, patterns within each group could be, particularly in commuter populations [16]. Nevertheless, future work should address this factor, not only by validating gender distribution but also including other demographic factors such as age and income. Finally, a potential line of research is the characterization of context surrounding perception, to answer questions such as: Is the *sexual* category completely related to harassment, or does vernacular language play a relevant role? What is the effect of interventions against harassment? These kinds of insights may be of value for practitioners and policymakers.

Acknowledgements. P.V. and E.G. were partially funded by Fondecyt de Iniciación project #11180913. All authors were partially funded by Concurso Interno de Investigación UDD #CI18.

REFERENCES

[1] Reeti Agarwal. 2008. Public transportation and customer satisfaction: the case of Indian railways. *Global Business Review* 9, 2 (2008), 257–272.

[2] Orna Blumen. 1994. Gender differences in the journey to work. *Urban geography* 15, 3 (1994), 223–245.

[3] Mairead Cantwell, Brian Caulfield, and Margaret O'Mahony. 2009. Examining the factors that impact public transport commuting satisfaction. *Journal of Public Transportation* 12, 2 (2009), 1.

[4] Yung-Hsiang Cheng. 2010. Exploring passenger anxiety associated with train travel. *Transportation* 37, 6 (2010), 875–896.

[5] Craig Collins, Samiul Hasan, and Satish V Ukkusuri. 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation* 16, 2 (2013), 2.

[6] Mariluz Congosto, Damaris Fuentes-Lorenzo, and Luis Sánchez. 2015. Microbloggers as sensors for public transport breakdowns. *IEEE Internet Computing* 19, 6 (2015), 18–25.

[7] Giovanni Costal, Laurie Pickup, and Vittorio Di Martino. 1988. Commuting—a further stress factor for working people: evidence from the European Community. *International archives of occupational and environmental health* 60, 5 (1988), 377–385.

[8] JM Del Castillo and FG Benitez. 2013. Determining a public transport satisfaction index from user surveys. *Transportmetrica A: Transport Science* 9, 8 (2013), 713–741.

[9] Eduardo Graells-Garrido, Diego Caro, and Denis Parra. 2018. Inferring modes of transportation using mobile phone data. *EPJ Data Science* 7, 1 (2018), 49.

[10] Joe B Hanna and John T Drea. 1998. Understanding and predicting passenger rail travel: an empirical study. *Transportation Journal* 38, 1 (1998), 38–46.

[11] GACCT Harman and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. *In ICWSM* (2014).

[12] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 237–246.

[13] Thorsten Hennig-Thurau, Caroline Wiertz, and Fabian Feldhaus. 2015. Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science* 43, 3 (2015), 375–394.

[14] Daehan Kwak, Daeyoung Kim, Ruilin Liu, Liviu Iftode, and Badri Nath. 2014. Tweeting traffic image reports on the road. In *2014 6th International Conference on Mobile Computing, Applications and Services (MobiCASE)*. IEEE, 40–48.

[15] Alexander Legrain, Naveen Eluru, and Ahmed M El-Geneidy. 2015. Am stressed, must travel: The relationship between mode choice and commuting stress. *Transportation research part F: traffic psychology and behaviour* 34 (2015), 141–151.

[16] Wendy Liu, Faiyaz Al Zamal, and Derek Ruths. 2012. Using social media to infer gender composition of commuter populations. In *Proceedings of the when the city meets the citizen workshop, the international conference on weblogs and social media*.

[17] Wendy Liu and Derek Ruths. 2013. What's in a Name? Using First Names as Features for Gender Inference in Twitter.. In *AAAI spring symposium: Analyzing microtext*, Vol. 13. 10–16.

[18] Anastasia Loukaitou-Sideris and Camille Fink. 2009. Addressing women's fear of victimization in transportation settings: A survey of US transit agencies. *Urban Affairs Review* 44, 4 (2009), 554–587.

[19] Gary Lynch and Susan Atkins. 1988. The influence of personal security fears on women's travel patterns. *Transportation* 15, 3 (1988), 257–277.

[20] Kelsey MacMillan and James D Wilson. 2017. Topic supervised non-negative matrix factorization. *arXiv preprint arXiv:1706.05084* (2017).

[21] Mehdi Menai et al. 2015. Walking and cycling for commuting, leisure and errands: relations with individual characteristics and leisure-time physical activity in a cross-sectional survey. *International Journal of Behavioral Nutrition and Physical Activity* 12, 1 (2015), 150.

[22] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45, 3 (2008), 211–236.

[23] Fabian Pedregosa et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[24] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[25] Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking gross community happiness from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, 965–968.

[26] Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies* 75 (2017), 197–211.

[27] Servicio Nacional de la Mujer. 2012. Estudio Acoso y Abuso Sexual en lugares públicos y medios de transporte colectivos. http://estudios.sernam.cl/documentos/?eMjI0MDIzOA==-Estudio_Acoso_y_Abuso_Sexual_en_lugares_publicos_y_medios_de_transporte_colectivos_. pp. 15.

[28] Thiago H Silva et al. 2013. Traffic condition is more than colored lines on a map: characterization of waze alerts. In *International Conference on Social Informatics*. Springer, 309–318.

[29] Patrick A Singleton. 2018. Walking (and cycling) to well-being: Modal and other determinants of subjective well-being during the commute. *Travel Behaviour and Society* (2018).

[30] Daphne Spain. 2000. Run, don't walk: how transportation complicates women's balancing act. In *Women's Travel Issues Second National ConferenceDrachman Institute of the University of Arizona; Morgan State University; Federal Highway Administration*.

[31] Evelyne St-Louis, Kevin Manaugh, Dea van Lierop, and Ahmed El-Geneidy. 2014. The happy commuter: A comparison of commuter satisfaction across modes. *Transportation research part F: traffic psychology and behaviour* 26 (2014), 160–170.

[32] Alois Stutzer and Bruno S Frey. 2008. Stress that doesn't pay: The commuting paradox. *Scandinavian Journal of Economics* 110, 2 (2008), 339–366.

[33] Universidad Alberto Hurtado, Observatorio Social. 2012. Encuesta Origen Destino Santiago 2012 (Informe Ejecutivo). http://www.sectra.gob.cl/biblioteca/detalle1.asp?mfn=3253. pp. 14, Accessed: 2018-09-24.