Portuguese speaking immigrant communities in Massachusetts: Assessing well-being through sentiment analysis of microblogging data

A thesis submitted by

Cara J. Foster-Karim

In partial fulfillment of the requirements for the degree of

Master of Arts

in

Urban and Environmental Policy and Planning

TUFTS UNIVERSITY

May 2015

Copyright 2015, Cara J. Foster-Karim

Advisor: Justin Hollander, Reader: Erin Graves

Abstract

Immigrant communities in Massachusetts Gateway Cities face a number of economic and social challenges that can be difficult to understand or quantify through traditional research methods. This thesis explores the use of sentiment analysis of microblogging data as an alternative method for assessing well-being of immigrant communities, with a focus on Portuguese speaking immigrants. I collected Tweets from four key cities in Massachusetts and analyzed them using two sentiment lexicons, one in Portuguese and one in English. I compared results between languages as well as correlated with a number of traditional indicators of well-being gathered from U.S. Census data. I found that the results from the English analysis were overall more positive than those from the Portuguese analysis, but most differences were not statistically significant. I also found some correlations between the demographic data and the sentiment analysis results with promising implications for further research.

Acknowledgements

I would first like to thank my Advisor, Justin Hollander, for consistent patience, enthusiasm, support, and always being willing to talk through my research framework and questions throughout this process. I would also like to thank my Reader, Erin Graves, for taking the time to share her expertise and provide comments and valuable insights into my research. I would especially like to thank my amazing husband, Tay Karim, for putting me through graduate school, for always being supportive and positive, letting me use his computer to collect data, providing technical support, and writing a new version of the sentiment analysis program that allowed me to analyze data much, much faster. I would also like to thank my mother, Sindia Foster, for gracious proofreading assistance. Finally, I am grateful to the rest of the Urban Attitudes Lab research team for their suggestions, technical support, and sharing of sources; especially Andrew Wiley and Dibyendu Das.

Table of Contents

Abstract **ii**

Acknowledgements **iii**

List of Tables **vi**

List of Figures **vii**

Chapter 1: Introduction **1**

Chapter 2: Literature Review **5**

Portuguese speaking immigrants in Massachusetts 5

Gateway Cities in Massachusetts 8

Assessing well-being 9

Microblog sentiment analysis 10

Uses of sentiment analysis 11

Microblog sentiment analysis in Portuguese 13

Urban applications of microblog sentiment analysis 14

Chapter 3: Methodology **16**

Data collection locations 18

Sentiment analysis 24

Statistical tests 27

Qualitative content analysis 28

Chapter 4: Results and Analysis **33**

Data collected 33

Significant results 37

Demographic variables 39

Demographic correlations 43

Results from qualitative content analysis 46

Chapter 5: Conclusions and recommendations 49

Revisiting research questions 49

Policy implications 51

Limitations 51

Recommendations for further research 54

Appendices 56

Appendix A: Map of Brockton data collection area 56

Appendix B: Map of Lowell data collection area 57

Appendix C: T-tests between proportions of positive Tweets in Portuguese results 58

Appendix D: T-tests between proportions of positive Tweets in English results 59

Appendix E: All results in Portuguese and English from all locations 60

Appendix F: Words removed from Portuguese lexicon 63

Appendix G: Correlation table 65

Appendix H: Demographic variables by Portuguese and Brazilian ancestry for each city 67

Appendix I: All Portuguese hand coding results 68

Appendix J: English hand coding results 71

Bibliography 73

List of Tables

Table 1: Data collection dates for each location 17

Table 2: Coordinates and areas of data collection locations 20

Table 3: Discrepancies between each sentiment lexicon used 27

Table 4: Method for qualitative content analysis 30

Table 5: Portuguese results for primary data collection areas 34

Table 6: English results for primary data collection locations 35

Table 7: Results from control locations 36

Table 8: Significance of differences between languages in each location 38

Table 9: Key demographic variables 40

Table 10: Demographic variables for residents of Portuguese or Brazilian ancestry 41

Table 11: Massachusetts city rankings based on results of sentiment analysis 44

Table 12: Massachusetts city rankings based on demographic variables 45

Table 13: Results from manual qualitative content analysis 48

List of Figures

Figure 1: Primary data collection locations in Massachusetts 19

Figure 2: Fall River data collection area 22

Figure 3: New Bedford data collection area 24

Figure 4: Example of Tweets analyzed by computer program, prior to applying qualitative content analysis 29

Chapter 1: Introduction

Immigrant communities in Massachusetts face a number of economic, social, and political challenges. In order to identify patterns and recommend and implement policy changes to be more inclusive of immigrants in Massachusetts, it is necessary to be able to assess and measure the well-being of these communities. Currently, census data, and economic indicators, as well as traditional surveys, interviews, and case studies are the most common and effective ways to study and assess these communities. However, these methods have their limitations including being slow and expensive to implement, and the fact that, especially with many undocumented immigrants living in Massachusetts, this population can be difficult to identify and study effectively. The purpose of this thesis is to propose and evaluate a method for supplementing traditional indicators of well-being, such as demographic and economic data, with sentiment analysis of microblogging data. This thesis will specifically focus on Portuguese speaking immigrant communities in Massachusetts Gateway Cities, using microblogging data collected from a few of these cities and analyzed for sentiment in Portuguese and English.

Gateway Cities in Massachusetts are specifically relevant to this study because they tend to have large immigrant populations (Immigrant Learning Center, 2012). “Gateway Cities” is a term used to identify cities that during the Industrial Revolution and into the early twentieth century served as

major destinations for immigrants to the United States, often because they could offer skilled labor to the textile mills or other industries located there. These cities are predominantly older industrial cities with decaying infrastructure and limited economic resources to assist newcomers. In 2008 the chief executives of eleven cities signed a Compact identifying these locations as "Gateway Cities", including: Fall River, New Bedford, Brockton, Fitchburg, Lawrence, Lowell, Haverhill, Springfield, Holyoke, and Pittsfield. The goal of this Compact is so that these cities can work together to lessen the economic gap between them and rest of the state (UMass Dartmouth 2014a). In 2010, the Massachusetts legislature officially codified the term (Foreman & Larson, 2014).

These post-industrial cities face a number of economic and social disadvantages. For example, 30% of Massachusetts residents living below the poverty line are in Gateway Cities, even though these cities account for only 15% of the total population (UMass Dartmouth, 2014b). According to Census data for New Bedford, a Gateway City which is also one of my study areas, 19.2% of the population is foreign born, and 60.2% of foreign born residents are limited English proficient (LEP). The median household income for New Bedford is $36,789 while for Massachusetts as a whole the median household income is $66,658 (United States Census Bureau, 2014c). Several of the Gateway Cities are also among the cities in Massachusetts with the highest numbers of immigrants (Immigrant Learning Center, 2012). In fact, the foreign-born population of many Gateway cities is trending back towards the

percentage of the early twentieth century (Foreman & Larson, 2014). This study hopes to examine the well-being of Portuguese speaking immigrants living in these Gateway Cities.

As a case study of the immigrant population of these Gateway Cities, my research will focus on Portuguese speaking residents. This is a diverse group, including people born in or with ancestry from Portugal, Cape Verde, and Brazil, along with smaller numbers from other Portuguese speaking countries (Massachusetts Alliance of Portuguese Speakers, 2008). Portuguese speakers as a group are especially relevant to focus on because Massachusetts has more Portuguese speakers than any other state, based on American Community Survey Data, 2006-2010 estimate (Massachusetts Alliance of Portuguese Speakers, 2012). The largest numbers of foreign-born Portuguese speakers in Massachusetts are originally from Brazil. In fact, in a 2012 report from the Immigrant Learning Center, which was based on 2009 data, Brazil was found to be the top country of origin for all foreign-born residents of Massachusetts (Immigrant Learning Center, 2012). Portuguese speaking immigrants also have had significant historical ties to Massachusetts, as they have been immigrating to Massachusetts since the 1800's (Bailey, 2000). While the country of origin has shifted--earlier Portuguese speaking immigrants were primarily from Portugal or the Azores, and today they are predominantly from Brazil--this group has had a long standing cultural impact on certain areas of Massachusetts. There is significant overlap between Gateway Cities and Portuguese speaking

immigrant communities, with several of the Gateway Cities among the cities that have the highest concentrations of Portuguese speakers, including: Fall River, New Bedford, Lowell, and Brockton (Massachusetts Alliance of Portuguese Speakers, 2008). These four cities will be the focus of my study.

I seek to contribute to the study of the well-being of immigrant communities in Massachusetts by assessing the well-being of Portuguese speaking immigrants in Gateway Cities using a sentiment analysis of microblogging data. In order to do so, I will collect geotagged microblogging data using the Twitter Streaming API from four key Gateway Cities in Massachusetts: Fall River, New Bedford, Lowell, and Brockton. I then will conduct a sentiment analysis in English and Portuguese of these collected microblogs using sentiment expressed through social media as an indicator of well-being. Finally, I will compare the results of the qualitative sentiment analysis with traditional quantitative indicators of economic and social well-being based on US Census data. My research questions include:

- How do Portuguese speaking microbloggers in Massachusetts Gateway Cities compare to English speaking microbloggers in terms of overall happiness and attitude?
- How do the results of a sentiment analysis of microblog data compare with traditional indicators of well-being in Massachusetts Gateway Cities?

Chapter 2: Literature Review

In order to pursue my primary research question is "How does microblog sentiment in Portuguese and English compare with traditional indicators of well-being for immigrants in Massachusetts Gateway Cities?" I will first consider how the various aspects of this topic have been addressed in the literature. This literature review will cover existing research on Portuguese immigrant communities in Massachusetts, economic indicators in Gateway Cities, well-being and how it is measured, and how sentiment analysis of microblog data has been used and compared with traditional forms of data collection in existing studies. Finally, I will look at applications of microblog sentiment analysis to other issues on the urban scale and consider the case of microblog sentiment analysis in a cross-cultural or multilingual context. Subsequent to the literature review I will introduce my analytical approach.

Portuguese speaking immigrants in Massachusetts

Immigrants from Portugal have a long history in Massachusetts. There are some records of a very few Portuguese settlers in New England dating back to the Revolutionary War (Pap, 1981) and larger numbers began to arrive in the early 1800s (Pap, 1981; Bailey, 2000). The first Portuguese speakers to settle in Massachusetts were sailors from the Azores, Madeira, and Cape Verde (islands which at the time all belonged to Portugal) picked up by captains of whaling ships from New Bedford in order to supplement their crews. This was an attractive career option for these sailors as it

allowed them avoid being drafted to join the military back in Portugal (Pap, 1981). According to Pap (1981) the whaling captains often exploited these sailors to encourage them to abandon ship so the captains would have an excuse to avoid paying them full wages. Once they had arrived in New England, many of these sailors settled in New Bedford and Fall River (Pap, 1981). The first record of a Portuguese name on the customs register in New Bedford dates from 1817, and in Fall River the first recorded Portuguese residents appear to have arrived in 1855 (Pap, 1981). Immigration from Portugal continued in greater numbers throughout the 19th century and the early 20th century, until Congress passed laws to limit immigration with the quota system in the 1920s. After a volcanic eruption in the Azores in 1957, the Azorean Refugee Acts of 1958 and 1960 were passed, beginning to open up opportunities for more Portuguese immigrants to come to the United States (Williams, 1982). Portuguese immigration increased through the 1960s and 1970s, both due to economic growth in the United States and reform of immigration laws, and political and economic problems in Portugal (Williams, 1982). Since, then, the numbers of immigrants from Portugal itself have been steadily decreasing (Bailey, 2000). More recent Portuguese speaking immigrants to Massachusetts have been from Cape Verde and especially from Brazil (Lima & Siqueira, 2007).

There is little research that I have been able to find on Portuguese speaking immigrants in Massachusetts as a group, outside of limited information provided by the Massachusetts Alliance of Portuguese Speakers.

Most studies either categorize Portuguese speakers with all other Latino immigrants, or they tend to focus on one national or ethnic group such as Lima and Siquiera's profile of Brazilians in the Massachusetts (2007). Siqueira and Jansen (2012) also have a study on working conditions for Brazilian immigrants in Massachusetts. A few older studies focus on Portuguese immigrants as an ethnic group, not including immigrants from other Portuguese speaking countries (Silvia, 1976; McGowan, 1976). A study by Jouet-Pastre and Braga looks at community-based learning in Portuguese speaking communities in Massachusetts, but it primarily evaluates a pedagogical technique and focuses very little attention on the actual communities in question (2005). The best community-wide information results from a Community Health Profile conducted by Massachusetts Alliance of Portuguese Speakers (2012). This study reviews a number of indicators such as education level and access to health insurance and makes comparisons between Portuguese speaking immigrants, other Latino immigrants, and Anglo-Americans. According to this Community Health Profile, there are 145,388 foreign born Portuguese speakers in Massachusetts, and 277,360 Massachusetts residents reporting ancestry from Portugal, Cape Verde, or Brazil. New Bedford and Fall River have the largest numbers of Portuguese speakers (14,544 and 12,753, respectively). Portugal is the primary country of origin for Portuguese speakers in New Bedford and Fall River. Brockton has 8,406 Portuguese speakers, with Cape Verde being the top country of origin. Brazilians are the largest group in

Lowell, although the total number of Portuguese speaking residents is only 4,419. These numbers are based on 2006-2010 American Community Survey data, 5-year estimates. Other key findings from the Community Health Profile include that Portuguese speakers are less likely to have health insurance than are Spanish or English speakers, and that Portuguese speakers are over three times as likely as English speakers to report themselves in poor or fair health (Massachusetts Alliance of Portuguese Speakers, 2012).

Gateway Cities in Massachusetts

While the term "Gateway City" may originally have referenced the impact of these cities in driving economic growth during the Industrial Revolution, the phrase also evokes the role of these cities both historically and today as an entry point for immigrants (Foreman & Larson, 2014). The Massachusetts Legislature has specified a formula to determine if a city is a "Gateway City" using both economic and demographic indicators (Foreman & Larson). The Urban Initiative at UMass Dartmouth has a number of reports on economic indicators in Gateway Cities along with proposals for improved economic development (2008). However, these proposals are more focused on plans for the future than on reporting on the current situation. Leavy-Sperounis (2010) also has a paper on green job creation in Gateway Cities. MassINC has also produced several reports on Gateway Cities in Massachusetts, most recently one focusing on immigrant entrepreneurship in Gateway Cities (Foreman & Larson, 2014). This report argues that in order to overcome their economic problems, Gateway Cities need to focus on

developing their strengths. The growing immigrant population in many of these cities is one of the most important strengths of Gateway Cities. Not only does the population represent a growing consumer base and labor force, but there is evidence that immigrants tend to start new businesses at a higher rate than native born residents, providing employment opportunities and a boost to the economy over all (Immigrant Learning Center, 2010).

Assessing well-being

Well-being may be a far more important way to assess the status of a community or society than the economic indicators that are usually used (Diener & Seligman, 2004). Diener and Seligman argue that well-being is much more predictive of worker productivity, mental and physical health, and social and community relationships than is economic status. However, it is somewhat imprecise to try to measure or even define something as subjective as well-being. Diener and Seligman (2004) define well-being as "peoples' positive evaluations of their lives, includes positive emotion, engagement, satisfaction, and meaning" (p.1). They point out that while economic factors do influence well-being, social relationships and physical health are have a greater impact (Diener & Seligman, 2004). Van Kamp, Leidelmeijer, Marsman, and De Hollander describe a similar concept, "quality of life", as the overlap of human community, natural environment, and economics (2003). On an individual scale, self-reported "well-being" can be measured with a survey or questionnaire, but this would be cost-prohibitive to administer on a large scale (Quercia, Ellis, Capra, & Crowcroft, 2012).

Fortunately, studies have found that results of sentiment analysis of social media content correlate strongly with self-reported life satisfaction for individuals (Kramer, 2010). This can also be applied to communities or even whole countries--Kramer used a sentiment analysis of Facebook posts to measure "Gross National Happiness" over time (2010). Quercia et al. took this a step further and found a strong relationship between overall sentiment detected from Twitter data and economic status at the community and neighborhood level (2012).

Microblog sentiment analysis

As an emerging field, analyzing microblog data for information on social issues has both strengths and weaknesses. It is a relatively fast and low cost method of collecting freely volunteered opinions in real time from a wide range of the public on a wide range of topics. This is much simpler, cheaper, and faster than conducting surveys or interviews, for example. However, there are also limitations to consider. Use of social media to express opinions and sentiment is much more pervasive among certain age groups or among those who have more access to smart phones and computers than it is among other groups. A Pew Research Center report from 2012 found that only 15% of adults in the United States used Twitter, and these were most likely to be between the ages of 18 and 29 and live in urban areas (Smith & Brenner, 2012). Mislove, Lehmann, Ahn, Onnela, and Rosenquist additionally found that Twitter users were significantly more likely to be male and live in densely populated areas (2011). There are

socioeconomic, linguistic, and cultural factors that may also impact use of social media. Thus, any social media or microblog data collected in this way cannot be said to be a random sample of the population to be studied, and it is important to be aware that key demographic groups may be under represented (Hollander, Graves, & Levanthal, 2014).

Uses of sentiment analysis

Sentiment analysis is a quasi-qualitative analytical method, a form of content analysis that can be applied to large data sets including social media data sets. Unlike traditional content analysis in which a researcher reads through a document and codes certain words and phrases, sentiment analysis is a more automated process, using a sentiment dictionary and a computer program to analyze large data sets.

Sentiment analysis of microblogging data has been used to consider social issues in a variety of studies. For example, sentiment analysis can be used to assess the public mood in response to events. Bollen, Mao, and Pepe (2011) conducted a Twitter sentiment analysis in which they considered nationwide sentiment over a six-month period in 2008. They calculated a daily mood for their entire pool of data, and correlated that with external events such as elections and holidays. Several other studies have compared microblogging sentiment analysis with the results of elections (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Gordon, 2013). Of special interest are those studies in which sentiment analysis has been used to

compare different geographic areas. For example, Quercia et al (2012) compared sentiment analysis of Tweets geotagged to different areas of London, and found a strong correlation between expressed positive sentiment and higher socioeconomic variables for each area. A number of other studies have also used geotagged tweets to look at differences between different geographic areas (Mearns et al., 2013; Mitchell, Frank, Harris, Dodds, & Danforth, 2014; Lovelace, Malleson, Harland, & Birkin, 2013; Antonelli et al., 2014; Balduini et al., 2013; Bertrand, Bialik, Virdee, Gros, & Bar-Yam, 2013).

My study will be particularly focused on comparing results between different languages. Several studies have used microblogging data to make comparisons between linguistic and cultural groups (Hale, 2014; Magdy, Ghanem, Musleh, & Mokbel, 2014; Mocanu et al., 2013; Wilkinson & Thelwall, 2012). Hale (2014) focuses on Twitter users who Tweet in more than one language to show how these users are instrumental in bridging language gaps and disseminating information from users of one language to users of another. Madgy et al. (2014) analyze geotagged Tweets from a number of countries around the world in order to compare linguistic diversity between different countries. Their unit of comparison is the country scale, and rather than analyzing the content of each Tweet they simply use the language tag associated with each Twitter user.

In a similar study, Mocanu et al. (2013) use language data associated with geotagged tweets to map language usage on Twitter both within and between countries. Wilkinson & Thelwell (2012), on the other hand, focus only on English but they compare how different topics are emphasized in different international and cultural contexts. Cui, Zhang, Liu, and Ma (2011) discuss the methodology of using emoticons to detect sentiment as way to bridge the gap between Tweets in different languages. Gao, Abel, Houben, and Yu (2012) compare sentiment analysis of microblogging data inside and outside of China by comparing results from Twitter and from Sina Weibo, the most popular Chinese microblogging service. Poblete, Garcia, Mendoza, and Jaimes (2011) compare Twitter sentiment and languages used in the top ten countries for Twitter usage worldwide.

Microblog sentiment analysis in Portuguese

According to Poblete et al., Brazil ranks second in the world for number of Twitter users (after the United States) and the ratio of Tweets per user is much higher than in the United States, making Portuguese one of the world's most Tweeted languages (2011). Thus it makes sense that there would be a number of studies that have focused on sentiment analysis of Tweets in Portuguese. Souza and Viera (2012) have produced an excellent review of the various sentiment lexicons available in Portuguese, as well as an analysis of the challenges of using sentiment analysis on Twitter data. De Araujo et al. (2012) apply sentiment analysis of Portuguese Tweets to public health topics. Santos, Esmin, Zambalde, and Nobre (2011) address marketing

and product development applications of sentiment analysis of Portuguese Tweets. A number of articles also address the different types of sentiment lexicons available to use on Portuguese Twitter data. Some of these focus on evaluating the effectiveness of one particular lexicon (Balage Filho, Pardo, & Aluiso, 2013; Gonçalves, Dores, & Benevenuto, 2012). Others, in a similar vein as Souza and Viera, attempt to give an overview of the various lexicons and methods available (Souza, Vieira, Busetti, Chishman, & Alves, 2011; Araújo, Gonçalves, & Benevenuto, 2013).

Urban applications of microblog sentiment analysis

My study falls within the category of urban applications of sentiment analysis. While the field is still new, Twitter sentiment analysis has been applied successfully to urban studies and urban planning topics. Antonelli et al. (2014) and Balduini et al. (2013) look at Twitter as a way to assess reactions to city-scale events, while MacEachren et al. (2011) apply similar methods to crisis management. Bertrand et al. (2013) apply sentiment analysis to Tweets from New York City to see how sentiment changes within different areas of the city and over time. Lovelace et al. (2014) consider a very small scale as they compare how many visitors frequent different museums in Yorkshire, England based on Tweets about the museums or Tweets sent from the geographic locations of the museums. Geotagged tweets have also been used track movement of people over time (Fujisaka, Lee, & Mumiya, 2010) and to determine land use in urban environments (Frias-Martinez, V., Soto, Hohwald, & Frias-Martinez, E., 2013). However, the

majority of these studies use only the quantitative data available from Twitter, rather than qualitatively analyzing the content of Tweets. One of the most comprehensive of these studies which does use sentiment analysis is that of Mitchell et al. (2013), which looks at happiness between States and urban areas with the United States and compares their sentiment analysis results to a number of other indicators of well-being, such as Gallup polls and gun violence rates. One of their more interesting findings was that areas with higher numbers of Tweets per capita tend to have less positive sentiment. They also correlated happiness from sentiment analysis with Census data and found a strong correlation between cities with a higher percentage of white, married, higher income residents and cities with higher happiness scores (Mitchell et al., 2013). By analyzing microblog data in different languages but from the same geographic area I hope to fill a gap in the literature by using sentiment analysis to make observations about how different groups of people experience the same space differently.

Hollander et al. (2014) produced a qualitative study of Twitter content to study attitudes relating to child and family policies and other urban issues within a specific urban area. This study is the closest in structure and scope to my research goals, and uses the specific qualitative sentiment analysis methods that I am using, applied to an urban planning-related topic. I will be basing my methods and approach heavily on their study.

Chapter 3: Methodology

The primary focus of this study is to compare the overall sentiment of microblogs in Portuguese and in English that are geotagged to Gateway Cities in Massachusetts. In order to collect this data, I pursued the following method, based on the work of Hollander, Graves, and Levanthal (2014). I focused on five collection areas: four Gateway Cities including Brockton, New Bedford, Fall River, Lowell, and a larger area of Eastern Massachusetts. I also collected data from a three other control locations for comparison purposes: Cambridge, Massachusetts; São Paulo, Brazil; and Lisbon, Portugal.

In order to collect microblog data, I used a computer program developed by the Urban Attitudes Lab research team[1] to collect data using the Twitter Streaming API. This program collects and indexes a continuous stream of publically available Tweets, totaling less than 1% of all Tweets generated at a given location (Hollander et al., 2014). Geotagged Tweets were collected from all five collection areas for several time periods in December 2014 and January 2015. While other studies based on collected Twitter data collected data continuously for extended periods of time--9 weeks (Gordon, 2013), 445 days (Lovelace et al., 2014) or six months (O'Connor et al., 2010)--these researchers were all using software that had

[1] The Urban Attitudes Lab is a research team affiliated with the Department of Urban and Environmental Policy and Planning at Tufts University and directed by Professor Justin Hollander. The Urban Attitudes Lab studies how Big Data can be used and analyzed to inform urban planning and policy. For more information, see the research team website at: http://sites.tufts.edu/ualab/

the ability to sort individual Tweets by the date the were collected so that it was possible to analyze the data over time. Due to the limited time frame of this project and the limited nature of the program being used, I only collected data over a two-month period, and it was not completely continuous. The timing of the breaks in data collection were in part due to my own schedule of when I was available to collect the data and in part due to when updates to the program were released, so that I had to download a new version and stop data collection while I installed the latest version of the program. Data was collected over slightly different time periods for the different locations as I adjusted my research project structure. Since my analysis is focused on comparisons between locations and languages, not over time, I consolidated all my data sets from each time period into one file per location. See Table 1 for details.

Table 1: Data collection dates for each location

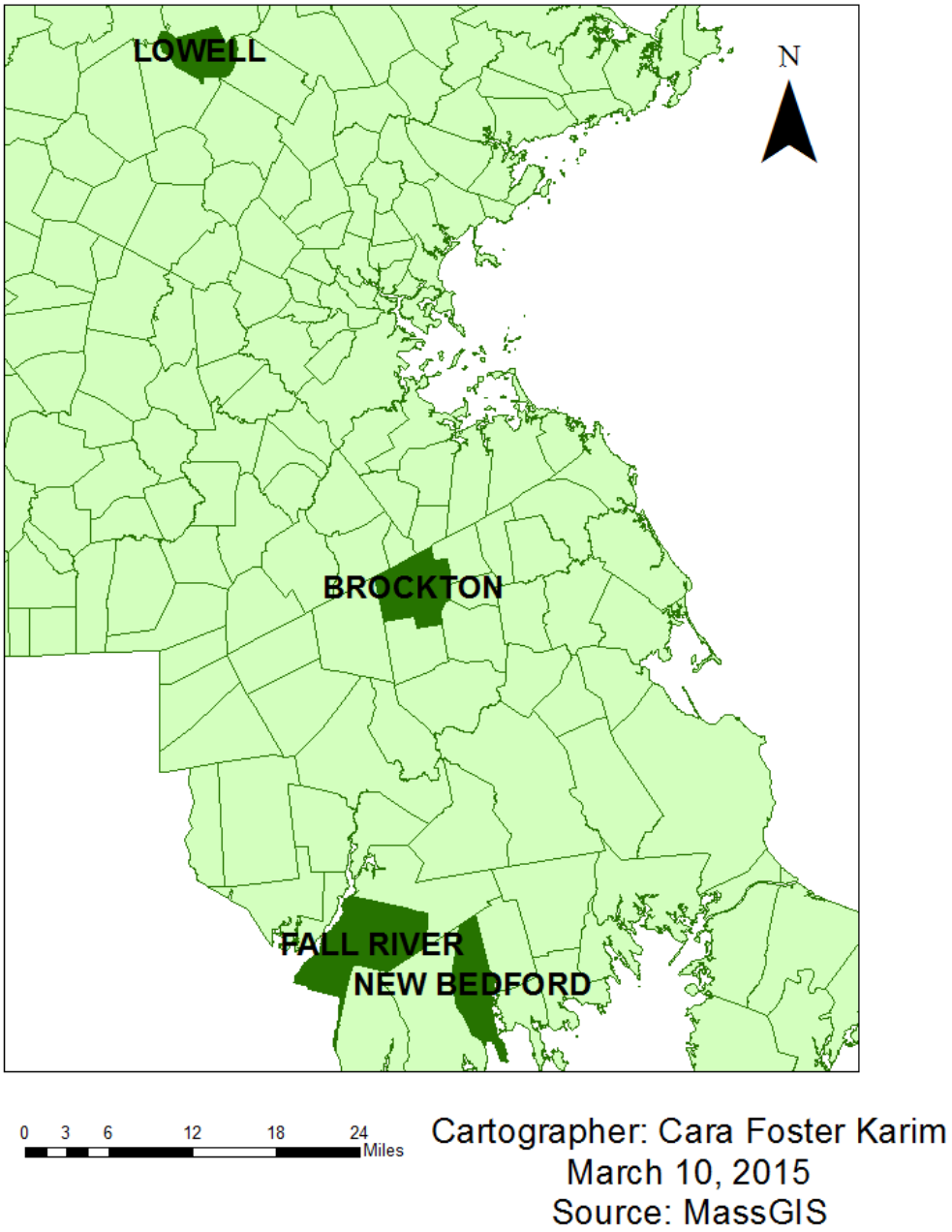
Location	Data collection dates
New Bedford	December 1-20th, January 6-19th, January 20-30th
Brockton	December 1-20th, January 6-19th, January 20-30th
Fall River	December 1-20th, January 6-19th, January 20-30th
Lowell	January 21st - February 16th
Eastern MA	December 1-20th, January 6-19th, January 20-30th
Harvard & MIT	January 26th - February 16th
São Paulo	January 20-30th
Lisbon	January 20-30th

The collected data was then stored in a .CSV format. Each individual Tweet has a unique ID number, a number assigned to the account holder, and a time and date stamp. Tweets collected from later versions of the program also have geographic coordinates, a city name, and the language listed on the account of the user who published that individual Tweet.

Data collection locations

To determine which cities to focus on, I compared the list of Gateway Cities from UMass Dartmouth with a list of cities with a high Portuguese speaking population, as reported by Massachusetts Alliance of Portuguese Speakers (MAPS). I chose four of the cities that are found on both of these lists: Fall River, New Bedford, Brockton, and Lowell. These cities are highlighted in Figure 1. I also collected data from a larger area covering the majority of eastern Massachusetts; excluding Cape Cod and the Islands. These five areas will be the focus of my data collection.

Figure 1: Primary data collection locations in Eastern Massachusetts



To contrast with these five locations, I also chose three control locations: São Paulo, Brazil; Lisbon, Portugal; and an area of Cambridge, Massachusetts containing Harvard and MIT's campuses. São Paulo and Brazil were chosen as locations that would be likely to generate microblogs mostly in Portuguese, which would both allow me to verify the reliability of the

Portuguese dictionary as well as consider differences in sentiment between Portuguese in areas where the language is a majority language versus a minority language. Cambridge was chosen as an area within Massachusetts that is known to be both very high income and have high rates of education, to contrast with the generally lower income Gateway Cities.

Table 2: Coordinates and areas of data collection locations

Location	Southwest Longitude	Southwest Latitude	Northeast Longitude	Northeast Latitude	Total Area of Data Collection
New Bedford	-70.977083	41.591415	-70.899655	41.745747	47 sq miles
Brockton	-71.100000	42.000000	-70.900000	42.400000	136 sq miles
Fall River	-71.200000	41.600000	-71.000000	41.800000	136.7 sq miles
Lowell	-71.400000	42.600000	-71.200000	42.700000	68.2 sq miles
Eastern MA	-71.338056	41.500000	-70.499722	42.799444	3,842.12 sq miles
Cambridge	-71.129803	42.354952	-71.079842	42.384766	5.29 sq miles
São Paulo	-46.899274	-23.750749	-46.300797	-23.451536	789 sq miles
Lisbon	-9.258021	38.692644	-9.075193	38.799608	72.5 sq miles

In order to direct the data collection, specific locations may be put into the computer program using latitude and longitude. For Fall River and Brockton, I collected Tweets from the smallest box I could get around each city boundary using a 0.1 degree (or 6 minute) grid because at the time when I started collecting data, the Urban Attitudes Lab believed that the Twitter API data was not accurate at any smaller scale, based on an answer received in an online forum about Twitter streaming by location (Nishikawa, 2014). To collect data for Fall River, for example, I used the northeast bounding

corner is at the coordinates N 41 degrees, 48 minutes (41.8) and W 71 degrees. The southwest bounding corner is at N 41 degrees and 36 minutes (41.6) and W 71 degrees and 12 minutes (71.2). As illustrated in Figure 2, this includes portions of several towns that surround Fall River.

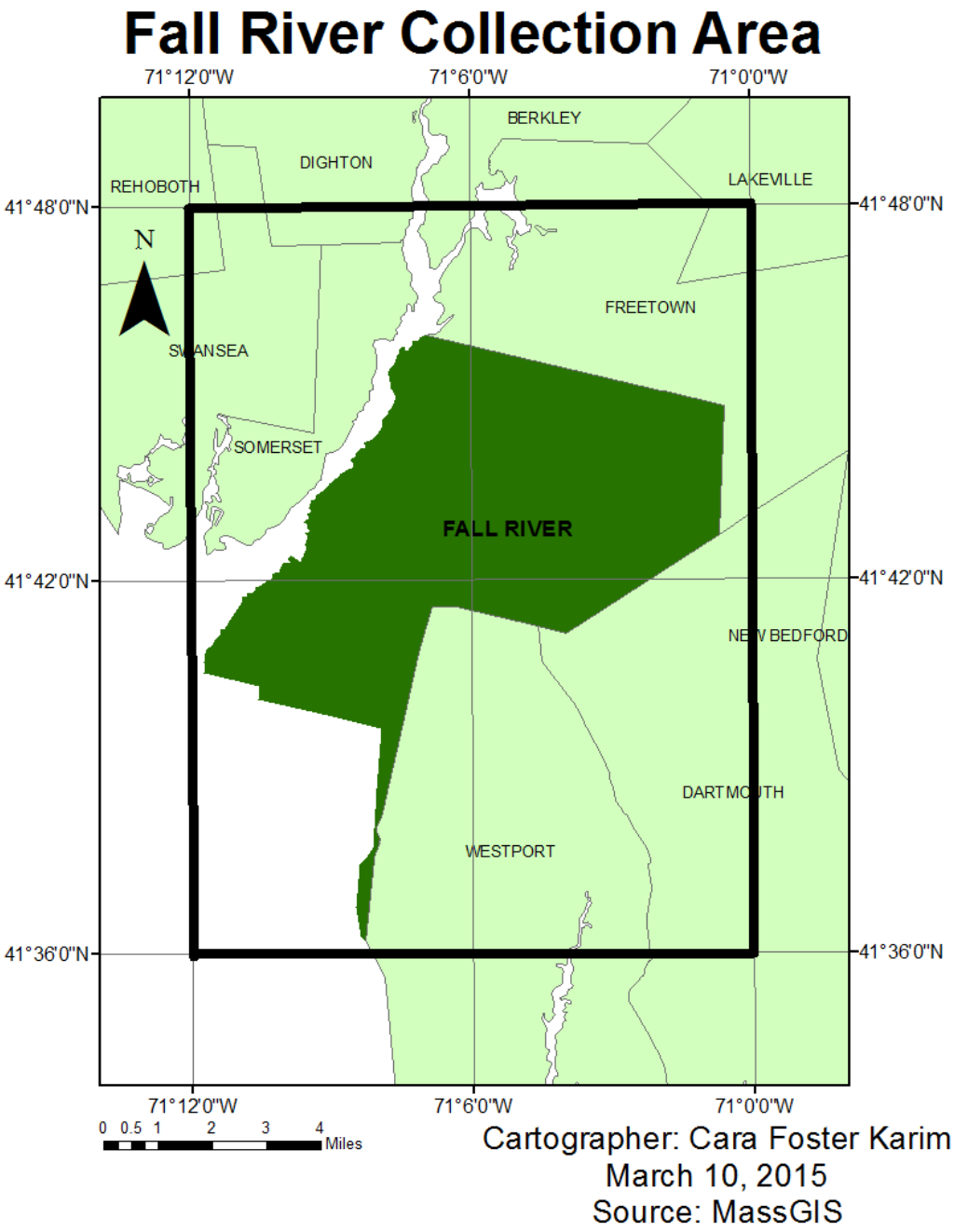


Figure 2: Fall River data collection area

While collecting Twitter data in this way is quite imprecise, this will still help to approximate the attitudes of Fall River residents. However, at a later point in the study my research team discovered that the Twitter Streaming API

was in fact available for smaller scales (Twitter, Inc., 2015). Thus my Eastern MA and New Bedford collection areas are rectangular boxes more closely fitted to the targeted area. For an example of this method, see Figure 3. Maps of my other primary data collection locations will be found in Appendices A and B.

Throughout my research on this project I used Tweets collected from within a geographic boundary as a proxy for residents of that area. Unfortunately, this does introduce increased error as Twitter users passing through a given area may skew the data. This is particularly something to be aware of in the context the Cambridge, MA control location as this area receives a large number of tourists and visitors. Other researchers, Quercia et al. in particular, have avoided this issue by identifying specific users by their user ID numbers and looking up the location indicated as a hometown in their public profile. However, tracking individual users was outside the scope of this project. Furthermore, this method is also not without issues as many users do not list a home location, and even if they do, this information may not be accurate or up to date.

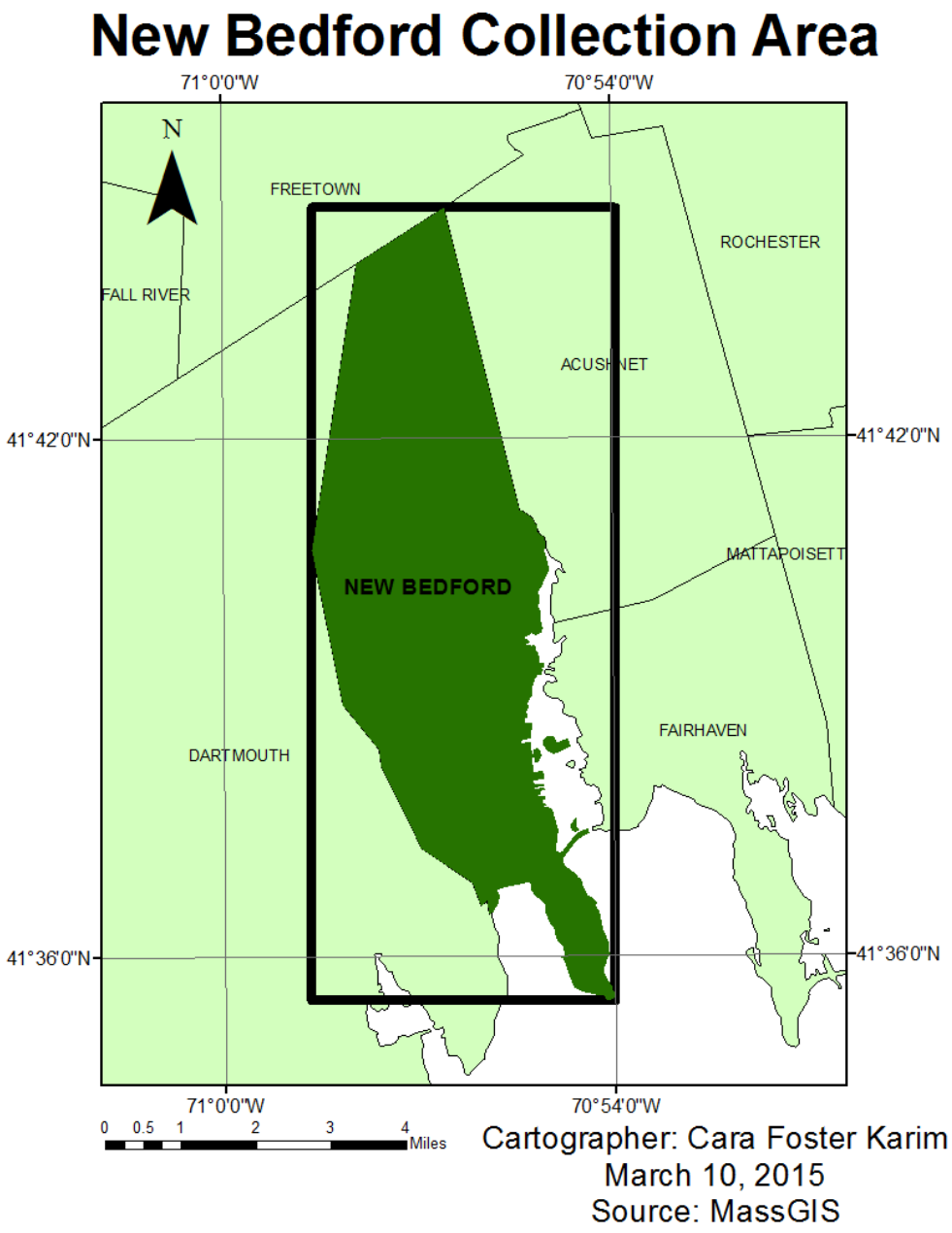


Figure 3: New Bedford data collection area

Sentiment analysis

Once I collected the stream of geotagged Tweets, I conducted the sentiment analysis. To do this I used the a computer program which takes as input a data set of Tweets and goes through the data set line by line to match

words in each Tweet with words in a sentiment dictionary and then assigns a score to each word that matches as well as an overall score for the data set. I used this program in combination with a sentiment analysis lexicon called SentiStrength (SentiStrength, 2015). I chose this lexicon because it is available in both English and Portuguese, so it would be possible to compare sentiment between Tweets in each language using the same scale. Furthermore, this lexicon is academically respected, having been evaluated in a number of peer-reviewed articles and has been used in multiple studies (Thelwall, Buckley, & Paltoglou, 2012; Pfitzner, Garas, & Scheitzer, 2012; Zheludev, Smith, & Aste, 2014; Kucuktunc, Cambazoglu, Weber, & Ferhatosmanoglu, 2012; Weber, Ukkonen, & Gionis, 2012; Garas, Garcia, Skowron, & Schweitzer, 2012; Grigore & Rosenkranz, 2011; Vural, Cambazoglu, & Senkul, 2012; Giannopoulos, Weber, Jaimes, & Sellis, 2012).

The SentiStrength sentiment lexicon is simply a list of common words that have been deemed to have sentiment, with a number value between -5 and +5 applied to each word, smaller values reflecting more negative emotions and positive values reflecting more positive ones. Two separate lexicons were installed, one for English and one for Portuguese. The results from the analysis using a lexicon take the form of a total tally of sentiment words, with the number of times each positive or negative word appears, which is then multiplied by each word's sentiment score to produce an overall sentiment score for each data set. In order to compare English and Portuguese language sentiment on the same set of data, the program is run

two different times, one with each lexicon, and the final scores are recorded. For the purposes of my research, I primarily used the total number of positive and negative Tweets from each data set, rather than using the score, although I did use the score for reference and comparison. In this way I was able to create a percentage value of overall positivity for each data set and made it possible to compare results between data sets of different sizes. With this method there is no need to separate English Tweets from Portuguese Tweets, because the program searches and scores Tweets word by word, rather than each Tweet as whole.

There are some important differences between the two lexicons that may have impacted the outcome of my research. The English sentiment lexicon contains 2476 words, while the Portuguese sentiment lexicon contains only 1964 words. Furthermore, the English lexicon contains 36% positive words and 64% negative words while the Portuguese lexicon contains 17% positive words and 83% negative words. I found upon running some preliminary analysis that the Portuguese dictionary also contained a large number of English or English slang words and terms used on the Internet or in text messages such as “lol”. While it may make sense in some contexts to include these words in a Portuguese sentiment dictionary because many Portuguese speaking Twitter users do in fact mix languages and use English terms mixed in with Tweets in Portuguese, this was unhelpful in the predominantly English context of Massachusetts because it caused the results to include a very large number of extraneous English

Tweets. In order to solve this problem, I manually removed 57 of the most frequently occurring English words from the Portuguese lexicon, resulting in a total list of 1907 words. See Appendix F for a full list of the words removed from the Portuguese dictionary. The new lexicon contains 16% positive words and 84% negative words.

Table 3: Discrepancies between each sentiment lexicon used

Language	Number of words	Positive words	Negative words
English lexicon	2476	36%	64%
Portuguese lexicon	1964	17%	83%
New Portuguese lexicon	1907	16%	84%

While this method is excellent for analyzing a large amount of data, there are several weaknesses to be considered. No sentiment lexicon can be exhaustive, and there may be words or phrases with obvious positive or negative meaning to a human reader that are not covered by the lexicon. Sentiment lexicons also have difficulty with sarcasm and context.

Statistical tests

I used a two sample T-test between percentages to see if the differences between the percentages of positive Tweets were significant across different locations and between each language at each location[2]. Then

[2] I used an online calculator from Social Science Statistics: http://www.socscistatistics.com/tests/Default.aspx

in order to see if there was any correlation between the results of the sentiment analysis and demographic and socioeconomic variables, I created a ranking system for each city being studied and a conducted bivariate ordinal correlation test to see how the variables related to each other.

Qualitative content analysis

Finally, in order to verify internal validity, I used qualitative content analysis to hand-code a sample of 300 of the collected and analyzed Tweets in each language for the Eastern Massachusetts collection location and compared my assessment to the results of the sentiment analysis from the computer program and the SentiStrength sentiment lexicon. In order to take a random sample, I created two .CSV files of just the Tweets that were identified as having sentiment from each dictionary. Then I used Excel to insert a column of random numbers into each file, sorted by the new column, and selected the first 300. Each row in the .CSV file contained the Tweet followed by the words that were determined to have sentiment and the score assigned each word. For an example of the format of the .CSV file that I started with, see Figure 4.

Figure 4: Example of Tweets analyzed by computer program, prior to applying qualitative content analysis

16	Finals got me like http://t.co/KDIDCwe6Tk	like	2						
17	@CreepyDemonLord // Ghira promise me you won't hurt him again...// *smiles*	smiles	2	won	3	promise	1	hurt	-2
18	@theblaze Never going to happen so who cares? Unfortunately.	cares	2						
19	@GenyFrumDaBlock bless ur soul lmao	bless	2	lmao	4				
20	@Fvck_AJ @Diamonx3 @MirandaaO @Gousby_617 nobody cares	cares	2						
21	@OmvrRivera: : You have such a clear face it's adorable asf but you never hmu thank you lol	clear	1	adorable	3	thank	2	lol	3
22	#RIPCaroline heaven gained a beautiful angel!! You're gonna be missed by so many!! Fly High beautiful	gained	2	beautiful	3	heaven	2	missed	-2
23	@g0dsbitch you win lmao	win	4	lmao	4				
24	I wanna go back to Old me and I won't tell nobody	won	3						
25	What's it like to stare into the eye of Satan's butthole	like	2						
26	And I also ordered riff raff tickets lmao	lmao	4						
27	@nicklicciardi12 got that right we gonna kill it doe.. love you a&f	love	3	kill	-3				
28	I have a massive crush on peter capaldi sorry nawt sorry	crush	-1	sorry	-1				
29	And you thought #riskmanagement jobs were hot. Hello #cybersecurity RT @kevinroose: Latest @Sony	dump	-1						
30	Cannot wait for 50 shades of grey to come out	grey	-1						
31	Some 80 year old lady just fist bumped me. How the fuck is your day going?	fuck	-4						
32	my work is gonna pay for me to take a baking class!!	pay	-1						
33	@ryan_palmer5: This is my bestfriend http://t.co/mtZw5SdVPv omg I love this//you	love	3						
34	@SHEEPISHLORDOFC LMFAOOO that baby is mad cute	cute	2	mad	-3				
35	Today would be a great day to find $20 on the ground	great	3						

I then added four columns; one for the word(s) that I determined to have sentiment, if different or in addition to the ones chosen by the computer analysis, one for the overall score that I would assign the Tweet, one column to indicate whether I agreed or disagreed with the sentiment dictionary's assessment, and one to indicate whether I shifted the score in a positive or negative direction. I also included a column to track what language the Tweet was in. From there, my basic methodology involved reading through the spreadsheet of Tweets one time to identify the language, and then reading it through a second time to score each Tweet. For comparison purposes, I also applied this method to a random sample of 300 Tweets from the analysis results using the older version of the Portuguese sentiment dictionary, before I had removed the English words. The following table (Table 4) highlights a small selection of English Tweets to illustrate my method. For example in the first Tweet listed, while the several positive words suggest that the sentiment is positive, the lexicon missed a key modifier, "RIP." If the

"RIPCaroline" hashtag is taken into account, the overall score of the Tweet is much less positive.

Tweet content	Sentiment words	Tweet score	Hand coding	New score
#RIPCaroline heaven gained a beautiful angel!! You're gonna be missed by so many!! Fly High beautiful	gained (2), beautiful (3), heaven (2), missed (-2)	5	"RIP" = -4	1
@g0dsbitch you win lmao	win (4), lmao (4)	8	"lmao" = 1	5
I wanna go back to Old me and I won't tell nobody	won (3)	3	"won't" =/ "won", "won't" = -1	-1
@nicklicciardi12 got that right we gonna kill it doe.. love you a&f	love (3), kill (-3)	0	"kill" in positive sense =2	5
I have a massive crush on peter capaldi sorry nawt sorry	crush (-1), sorry (-1)	-2	"crush" in positive sense = 2	1
Cannot wait for 50 shades of grey to come out	grey (-1)	-1	"cannot wait" = 3	2
@SHEEPISHLORDOFC LMFAOOO that baby is mad cute	cute (3), mad (-3)	-1	"mad" as intensifier = 1	3

Table 4: Method for qualitative content analysis

While I will address the majority of my findings from this qualitative content analysis in the results section, it is important to note that for the Portuguese sample, less than 20% of the Tweets identified as having Portuguese sentiment by the lexicon were actually in Portuguese. Another 20% were in Spanish and the rest were in English and other languages. For the English sample, however, over 93% of the results were actually in English. It is important to take into account these results of the qualitative analysis while considering the results from the quantitative analysis.

Chapter 4: Results and Analysis

In this section, I will first present the initial results from all of the data collection locations. Next I will discuss the statistical significance of differences in the results between different languages and between the data collection locations. Then I will present demographic information based on United States Census data for the Massachusetts collection locations and discuss some correlations between the demographic data the results of my analysis. Finally I will discuss the results of the qualitative content analysis I conducted.

Data collected

In total across all my data sets, I collected and analyzed 1,167,176 Tweets. Some of my collection areas overlapped, so not counting any of the potentially duplicated Tweets I had at minimum 814,315 discrete Tweets. Since the Twitter Streaming API releases a random sample of less than 1% of all public Tweets from any given geographic area it possible that there was in fact very little overlap in my data even for areas that overlapped geographically.

After running the analysis with the Portuguese sentiment lexicon on my primary data collection areas over all time periods, I was able to generate the following results.

Table 5: Portuguese results for primary data collection areas

Location	Total Tweets	Tweets w/ sentiment	Coverage	% Positive	% Negative	Sentiment score	Normalized score
Eastern MA	315448	4277	1.3%	50.8%	49.2%	-959	-0.2146
Fall River	30770	217	0.7%	50.7%	49.3%	-26	-0.1176
Brockton	273854	4211	1.5%	50.9%	49.0%	-958	-0.2184
New Bedford	37732	249	0.6%	50.8%	49.2%	-48	-0.1905
Lowell	38663	350	0.9%	50.8%	49.2%	-89	-0.2445

The Portuguese dictionary was only able to identify Tweets with sentiment in Portuguese in .66% to 1.54% of all Tweets ("coverage"). Of those, just over 50% were positive in all the locations. The "percent positive" value is calculated by dividing the number of positive Tweets by the number of positive and negative Tweets combined. The denominator might be slightly larger than the number of Tweets with sentiment listed in the third column because some Tweets have both positive and negative words, in which case they might be counted twice. The "overall score" is the sum of all the sentiment words identified by the lexicon, multiplied by the value or degree of positivity or negativity as measured by the sentiment lexicon. This value is hard to compare between locations because it is proportional to the number of words with sentiment in a data set, and each of my data sets is a different size, due to the length of time Tweets were collected, the size of the collection area, and other factors. To normalize it, I divided the score by the

number of positive and negative Tweets in each data set, producing the number displayed in the last column.

Table 6: English results for primary data collection locations

Location	Total Tweets	Tweets w/ sentiment	Coverage	% Positive	% Negative	Sentiment score	Normalized score
Eastern MA	315448	169086	53.6%	56.5%	43.4%	61200	0.3038
Fall River	30770	17895	58.2%	54.9%	45.2%	1570	0.0730
Brockton	273854	144171	52.6%	56.5%	43.5%	48358	0.2826
New Bedford	37732	21783	57.7%	53.9%	46.1%	-1717	-0.0649
Lowell	38663	22037	57.0%	54.6%	45.4%	2561	0.0968

The first major difference to note between these English results and the earlier Portuguese results are that the coverage is much higher, between 52% and 58% across all the locations. This reflects both that Massachusetts is an English majority area and the larger number of words in the English sentiment dictionary. The second major thing to note is that the percentage of positive Tweets is also higher in each collection area.

The following table shows my results from the three control locations, Lisbon, São Paulo, and the Harvard and MIT area of Cambridge.

Location	Language	Total Tweets	Tweets w/ sentiment	Coverage	% Positive	% Negative	Sentiment score	Normalized score
Lisbon	Portuguese	22156	7774	35.1%	55.0%	45.0%	-452	-0.0521
São Paulo	Portuguese	438048	154920	35.4%	56.7%	43.3%	26723	0.1514
Harvard and MIT	Portuguese	10505	134	1.3%	44.4%	55.6%	-44	-0.3259
Lisbon	English	22156	2797	12.6%	45.3%	54.7%	1297	0.4355
São Paulo	English	438048	63966	14.6%	48.9%	51.1%	60408	0.8937
Harvard and MIT	English	10505	5533	52.7%	62.2%	37.8%	5504	0.8508

Table 7: Results from control locations

Note that in Lisbon and São Paulo, the coverage percentage in Portuguese is much higher than in the Massachusetts cities, indicating a much larger pool of Tweets in Portuguese. However it is still not as high as the coverage in English in Massachusetts, probably due to limitations of the sentiment lexicon. Interestingly, the Cambridge collection area had the highest percentage of positive Tweets in English, but the lowest in Portuguese. If measuring by the normalized score, however, the English Tweets from São Paulo were more positive in sentiment than those from Cambridge. Also, the Lisbon and São Paulo locations had an even higher percentage of positive Tweets overall than the Massachusetts locations when measured in Portuguese, but less positive when measured in English.

Significant results

In order to determine which of these interesting differences were statistically significant, I ran a series of two-sample T-tests between percentages. I considered differences in the same language between each location, and then looked at differences between the results in each language at each location. I found that in both English and Portuguese the results were significantly different between most of the Gateway Cities and most of the control locations. When comparing the Gateway Cities with each other, however, there was no statistically significant difference in Portuguese positivity. In the English results, there was no significant difference between Lowell and Fall River and New Bedford, but Brockton was significantly different than the others. With the exception of the Brockton English results,

these findings validate my decision to group these four cities together for analysis purposes, since the differences between them in terms of sentiment scores are for the most part not statistically significant. See Appendices C and D for full tables of relationships between the results from all the data collection areas in each language.

I also looked at each location to see if there was a significant difference between the Portuguese and English positivity scores for a given location. These results are displayed in the following table.

Table 8: Significance of differences between languages in each location

Location	Portuguese positivity	English positivity	Z-score	p-value
Eastern MA	50.8%	56.5%	7.6716	0.0000*
Brockton	50.7%	54.9%	1.2535	0.1056
Fall River	50.9%	56.5%	7.3034	0.0000*
New Bedford	50.8%	53.9%	0.9963	0.1587
Lowell	50.8%	54.6%	1.4444	0.0749
Lisbon	55.0%	45.3%	-9.1318	0.0000*
São Paulo	56.7%	48.9%	-34.6467	0.0000*
Harvard and MIT	44.4%	62.2%	4.2122	0.0000*
Average for all four Gateway Cities	**50.8%**	**55.0%**	**-7.1297**	**0.0000***

* *Indicates significance at p< 0.05*

I found that the difference between the Portuguese and English percentages of positivity were statistically significant for the overall region of Eastern Massachusetts, Fall River, Lisbon, São Paulo, and Cambridge. In Eastern Massachusetts, Fall River, and Cambridge, the English results had a significantly higher percentage of positive Tweets. In Lisbon and São Paulo, however, the percentage of positive Tweets in English was significantly lower than those in Portuguese. I also aggregated the results for the four

Gateway Cities and I found the difference between their average scores to also be significant, with results in English significantly more positive than results in Portuguese.

Demographic variables

To further explore potential meaning in these results, I added demographic variables from the United States Census Quickfacts pages for each of the Massachusetts cities and accessed some additional data from the American Community Survey 5-year estimates (2009-2013 and 2006-2010). Please note that the US Census boundaries for each of these cities are not the same as those I have used for my data collection areas. For the most part my collection areas were larger than the official city boundaries, including areas from cities around each target city. For the Cambridge location, because I was trying to focus on the Harvard and MIT area, only about two-thirds of the area of the city is included in my collection area, although this two-thirds does include the most densely populated area of the city. These demographic variables are not intended to exactly represent the population of the data collection areas, but rather provide some background and a rough approximation.

Table 9: Key demographic variables

Demographic variable	Brockton	Fall River	New Bedford	Lowell	Average of Gateway Cities	Cambridge	Massachusetts overall
Population	94,089	88,697	95,078	108,861	96,681	107,289	6,745,408
% under 18	25.7%	21.5%	23.20%	23.7%	23.7%	11.4%	20.8%
% 65 and over	11.9%	15.1%	14.6%	10.1%	13.0%	9.5%	14.8%
% Black/African American	31.2%	3.9%	6.4%	6.8%	12.0%	11.7%	8.1%
% Hispanic	10.0%	7.4%	16.7%	17.3%	12.7%	7.6%	10.5%
% foreign born	24.7%	19.0%	19.9%	24.8%	22.2%	27.7%	15.0%
Median household income	$49,025	$33,211	$35,999	$49,452	$41,922	$72,529	$66,866
% with high school diploma	80.5%	70.3%	70.5%	78.8%	75.0%	93.7%	89.4%
Language other than English spoken at home	36.6%	34.0%	37.0%	41.9%	37.4%	32.1%	21.9%
Population density	4,398.40	2,681.90	4,754.30	7,842.10	4,919.00	16,470.20	839.4
Portuguese Ancestry	3,946	41,443	38,603	5308	22,325	2,208	311,767
Brazilian Ancestry	1,238	1,197	997	2901	1,583	(no data)	65,170
% of population with Brazilian or Portuguese ancestry	5.5%	48.1%	41.6%	7.5%	24.7%	2.0%	5.6%

Sources: Decennial Census and American Community Survey 5-year estimates, 2006-2010 and 2009-2013(U.S. Census Bureau 2014a, 2014b, 2015a, 2015b, 2015c, 2015d, 2015e)

These variables were chosen because they are often used as indicators of socioeconomic status or well-being, especially median household income or percent of adult population with a high school diploma. I chose percent foreign born and language other than English spoken at home as a way to measure the immigrant communities in these cities. Unfortunately data on Portuguese speakers is not available at the city level but there is American Community Survey data on people who reported Portuguese or Brazilian ancestry. I did not include the data compiled by the Massachusetts Alliance of Portuguese Speakers on the number of Portuguese speakers per city, as this was a composite result that they estimated from a combination of different United States Census sources.

Some important variables to notice from this table include median household income, population under 18 and 65 and over, high school graduation rates, and percent of residents who are foreign born. The four Gateway Cities studied have a lower median household income, a younger population, fewer high school diplomas, and higher percentage of foreign born residents than is true for Massachusetts as a whole. Furthermore, the proportion of residents who claimed Brazilian or Portuguese ancestry, 24.7%, is much higher than the statewide percentage of 5.6%.

The following table (Table 10) highlights some of these key variables for the segment of the population that indicated Brazilian or Portuguese ancestry in the four Gateway Cities. Brazilian or Portuguese ancestry is a

loose approximation for Portuguese speakers, as many residents may be of Portuguese or Brazilian heritage but not currently speak Portuguese. This data also ignores any residents of Cape Verdean background, as they are not tracked as a specific ancestry group by the Census. Nonetheless, Brazilian and Portuguese ancestry is the best proxy for Portuguese speakers available in the Census data.

Table 10: Demographic variables for residents of Portuguese or Brazilian ancestry

Demographic variable	Average for Brazilian and Portuguese ancestry	Average for four Gateway Cities
% with high school diploma or higher	70.4%	75.0%
Median household income	46,155	41,922
% under 18	24.1%	23.7%
% 65 and over	7.0%	13.0%
% Foreign born	53.8%	22.2%
Population	23,908	96,681

Sources: *United States Census Bureau (2015f, 2015g, 2015h, 2015i, 2015j)*

As you can see, while representing almost a quarter of the population of these Gateway Cities, there are some key differences between residents of Portuguese or Brazilian background and the rest of the population. The number residents of Portuguese or Brazilian ancestry who are under 18 is higher than average, while number who are 65 or over is much lower. This group is also almost twice as likely to be foreign born. They are somewhat less likely to have a high school diploma, although this differed between

those of Portuguese and Brazilian heritage—Brazilians had a higher than average high school graduation rate, and Portuguese had a lower than average high school graduation rate. To see all these demographic variables broken down between the two groups, see Appendix H.

Demographic correlations

In order to assess any correlations between my results and the demographic data, I first ranked each of the five Massachusetts collection areas on select demographic variables as well as how they performed in the results of the sentiment analysis. My advisor, Justin Hollander, ran a Spearman's rho test on these ranked cities to look at correlations between the sentiment variables and the demographic variables. I chose percent of positive and negative Tweets in Portuguese and English, the overall score for each language, and the normalized score for each language as my sentiment variables. For demographic variables, I chose population, percent foreign born, percent with a high school diploma or higher, median income, percent foreign born, percent under 18 and over 65, population density, and Portuguese and Brazilian ancestry. The rankings are displayed in the following two tables.

Table 11: Massachusetts city rankings based on results of sentiment analysis

City	Percent positive		Percent negative		Raw sentiment score		Normalized score	
Language	English	Portuguese	English	Portuguese	English	Portuguese	English	Portuguese
Brockton	2	1	4	5	1	5	2	3
Fall River	3	3	3	2	4	1	4	1
New Bedford	5	4	1	3	3	3	5	2
Lowell	4	2	2	4	2	4	3	4
Cambridge	1	5	5	1	5	2	1	5

Table 12: Massachusetts city rankings based on demographic variables

City	Population	Foreign born	Median income	High school education	Under 18	65 and over	Population density	Portuguese ancestry	Brazilian ancestry
Brockton	4	3	3	2	1	3	4	4	2
Fall River	5	5	5	5	4	1	5	2	3
New Bedford	3	4	4	4	3	2	3	1	4
Lowell	1	2	2	3	2	4	2	3	1
Cambridge	2	1	1	1	5	5	1	5	5

I found statistically significant ($p < .05$) correlations between some of the demographic and sentiment variables. Having a high population under 18 correlated positively with positive Portuguese sentiment scores, whereas a high population of residents with Portuguese ancestry correlated negatively with positive English sentiment scores. Conversely, a high population of residents with Brazilian ancestry correlated positively with positive Portuguese sentiment scores, which is expected in the light of the more positive Portuguese sentiment scores in the control locations with higher numbers of Portuguese speakers (São Paulo and Lisbon). A high percentage of residents who had at least completed high school correlated positively with a higher normalized sentiment score in English, and negatively with a higher normalized sentiment score in Portuguese. See Appendix G for the full table of all correlations between variables.

Results from qualitative content analysis

My traditional qualitative content analysis was valuable in that it verified the basic reliability of the sentiment lexicons for some purposes; it also highlighted many of the sentiment lexicons' weaknesses. On the positive side, of Tweets that were actually in the correct language, 81.1% of Portuguese Tweets and 88.5% of the English Tweets were correct in terms of the overall positivity or negativity of the sentiment score assigned by the computer program. However, for the Portuguese results, less than 20% of the Tweets flagged as having sentiment by the computer program were actually in Portuguese. Another 20% of those flagged as being Portuguese were

actually in Spanish. There are a number of factors impacting this poor data quality, besides the smaller number of words in the Portuguese lexicon. Due to the many cognates between Portuguese, Spanish, and English, it is impossible to eliminate all the erroneous Tweets. Many of the Tweets erroneously identified as Portuguese contained legitimate "Portuguese" words—they just also happen to be words in Spanish or English. For example, words such as "favor" and "original" have identical spellings and similar meanings across all three languages. The Portuguese lexicon may also have missed many Tweets that actually were in Portuguese, for a couple of reasons. In Portuguese, nouns have gender and can be masculine or feminine. Adjectives change form depending on the noun they are modifying. However the SentiStrength lexicon in Portuguese seemed to contain only the masculine form for most adjectives. The lexicon also left out common Portuguese internet abbreviations, such as "tb" for *também* (also) or "pq" for *por qué* (why or because). Finally, accents are important to correct Portuguese spelling, and words in the SentiStrength lexicon have accents. However the format of the Twitter data collected by the computer program did not allow for accents, meaning that if either the original Twitter user misspelled words without accents (which is common in informal settings or when typing on a cell phone), or they spelled them correctly and the computer program did not include them, the Portuguese lexicon would be unable to match those words. See Appendices I and J for a full list of hand-coded Tweets in each language.

Table 13: Results from manual qualitative content analysis

Language	% of Tweets in correct language	% of incorrect scores	% of correct scores	No. shifted in a positive direction	No. shifted in a negative direction	Original score for sample	Adjusted score for sample
Portuguese	19.3%	18.9%	81.1%	6	5	-91	-89
English	98.3%	11.5%	88.5%	8	26	151	96

Chapter 5: Conclusions and recommendations

In this concluding section I will consider my results in the light of my research questions as well as discussing potential policy implications of my results. I will also discuss the limitations of this study and propose some recommendations for future research.

Revisiting research questions

To assess how successful my study was, I will return to my research questions as a means of evaluation. My research questions were, "How do Portuguese speaking microbloggers in Massachusetts Gateway Cities compare to English speaking microblogggers in terms of overall happiness and attitude?" and "How do the results of microblog sentiment analysis compare with traditional indicators of well-being in Massachusetts Gateway Cities?"

I was able to answer each of these questions partially. Regarding the first research question, there was a measurable difference in the percent of positive Tweets measured in Portuguese and English in each location, with results of the English sentiment analysis coming up consistently more positive than the results of the Portuguese sentiment analysis in the Massachusetts Gateway Cities. However, this difference was only statistically significant for one of the Gateway Cities, Fall River. It was also statistically significant in in Eastern Massachusetts over all and in my control location of Cambridge. When I averaged the results from the four Gateway Cities, the

average Portuguese result was significantly less positive than the English result. I can answer my research question and observe from my results that the results of the English sentiment analysis were more positive than the results of the Portuguese sentiment analysis. However, because of the limited significance of these results and many limitations of the sentiment lexicon used, it is not possible to make any policy conclusions from these results.

In order to address my second research question, I looked at the correlation between a number of traditional indicators of well-being and economic success as measured by the United States Census Bureau, including American Community Survey data, and the results of my analysis. There was much less correlation here than I had anticipated. A statistically significant correlation only appeared for a few variables. I found that higher percentages of positivity in Portuguese and higher normalized sentiment scores in Portuguese correlated positively with younger populations and lower adult rates of high school graduation. Larger numbers of residents with Brazilian ancestry also correlated positively with more positive Portuguese sentiment, and larger numbers of residents with Portuguese ancestry correlated negatively with more positive English sentiment. While these results are interesting, it is clear that the sentiment analysis does not correlate in an obvious way with more common indicators such as median income. And since I cannot be sure what percentage of people Tweeting in a particular geographic area are actually residents of that area, these correlations may be less meaningful.

Thus, the answers to my research questions were somewhat inconclusive. However it is still clear that there is a difference in overall sentiment between Portuguese and English microbloggers in Massachusetts Gateway Cities, even if this difference is not statistically significant in most cases. Furthermore, while median income did not have the relationship to sentiment scores that I expected, the relationships between age, ancestry, and education levels with sentiment scores have interesting implications and are worthy of further study.

Policy implications

The limited number of statistically significant results from this study, as well as the many limitations I encountered in my research, mean that it is problematic to clearly identify specific policy implications. While many economic and social challenges face Portuguese speaking residents of Gateway Cities, it is not clear from the evidence that these challenges are more severe or different in nature than the challenges faced by other residents of these locations. Further studies are needed to fully address these questions.

Limitations

The many limitations of this study prevent me from making policy recommendations. The limitations fall into three main categories: the shortcomings of using Twitter itself as a research tool, the mistakes and

inconsistencies of my research methods, and flaws I discovered in the sentiment lexicons that I used for the analysis.

First, it is important to be aware of the inherent limitations of using microblog data for research purposes. Microbloggers are by no means a representative sample of the population of a geographic area, as previously discussed. The tiny fraction of all Tweets that are public, geotagged, and available from the Twitter Streaming API are even more limited. Furthermore, microblog data is by nature extremely poor in quality. Many of the Tweets collected for this study are promotional ones from companies or organizations, or multiple re-Tweets of the same content. There is also no easy way to tell, based on the data I collected, whether the Twitter user who posted a given message is actually a resident of the study area or simply passing through. It would be possible to look up the user ID of each user to see where they had indicated as their hometown--but not everyone lists this information, and people frequently move and fail to update it. It is also important to note that at best only about 50% of the Tweets collected had any sentiment at all (as measured by the lexicon).

There were also a number of limitations to my data collection and analysis methods. I encountered some technical difficulties with the program I used that meant that my data collection was often interrupted and was not constant. Also, due to the change in information I had about the scale at which Twitter streaming API data was available, some of the collection areas

were much larger, including more portions of other cities and towns, than others. Even for the areas where I did use the same method to delineate a collection area, geographic boundaries of cities varied in size. Also, due in part to technical issues and partly due to adaptations of my research plan, I did not collect data from all locations for the same period or length of time. There were also some inconsistencies that arose from issues with the computer program I used. As various updates were released over the course of my research, some of my later data sets included new information, such as the geographic coordinates of the Tweet location and the language that the user had specified on their account. But because this data was not available for all the versions, I was not able to use it in my analysis in a significant way.

Finally, there were a number of limitations that arose from the sentiment lexicons I used to analyze the collected data. I used two sentiment lexicons, one in Portuguese and one in English, which were designed by the same research team. The purpose of using the SentiStrength lexicon was that it was one of the only ones I was able to find that used the same scoring system in each language and had been evaluated for comprehensiveness and accuracy by academic sources. However once I started to use it, I found that there were major differences and inconsistencies between the two lexicons, making it very hard to reliably compare results between them. First, the Portuguese dictionary contained far fewer words than the English one, and a much larger percentage of those words were scored as negative. This would potentially color the results by reducing the percentage of Tweets that were

found to have sentiment in Portuguese, and making it more likely to identify more negative Tweets. Second, the Portuguese lexicon included many English words, particularly slang words or abbreviations common in internet usage such as "lol". While it is probably true that Portuguese speaking Twitter users commonly mix such English words into their Portuguese Tweets, when applying the Portuguese sentiment lexicon to a predominantly English data set, a very high percentage of extraneous Tweets were included in the analysis even though they were not actually in Portuguese at all. With both of the lexicons, I found through my qualitative content analysis that many common words were scored in a way did not reflect their most common usage. For example in the English lexicon, I found that the word "like" was given a fairly positive score, even though the vast majority of the time it appeared in the results, the actual meaning involved either making a comparison or simply using the word as filler, so that the overall content of the Tweet was often not positive at all.

Recommendations for further research

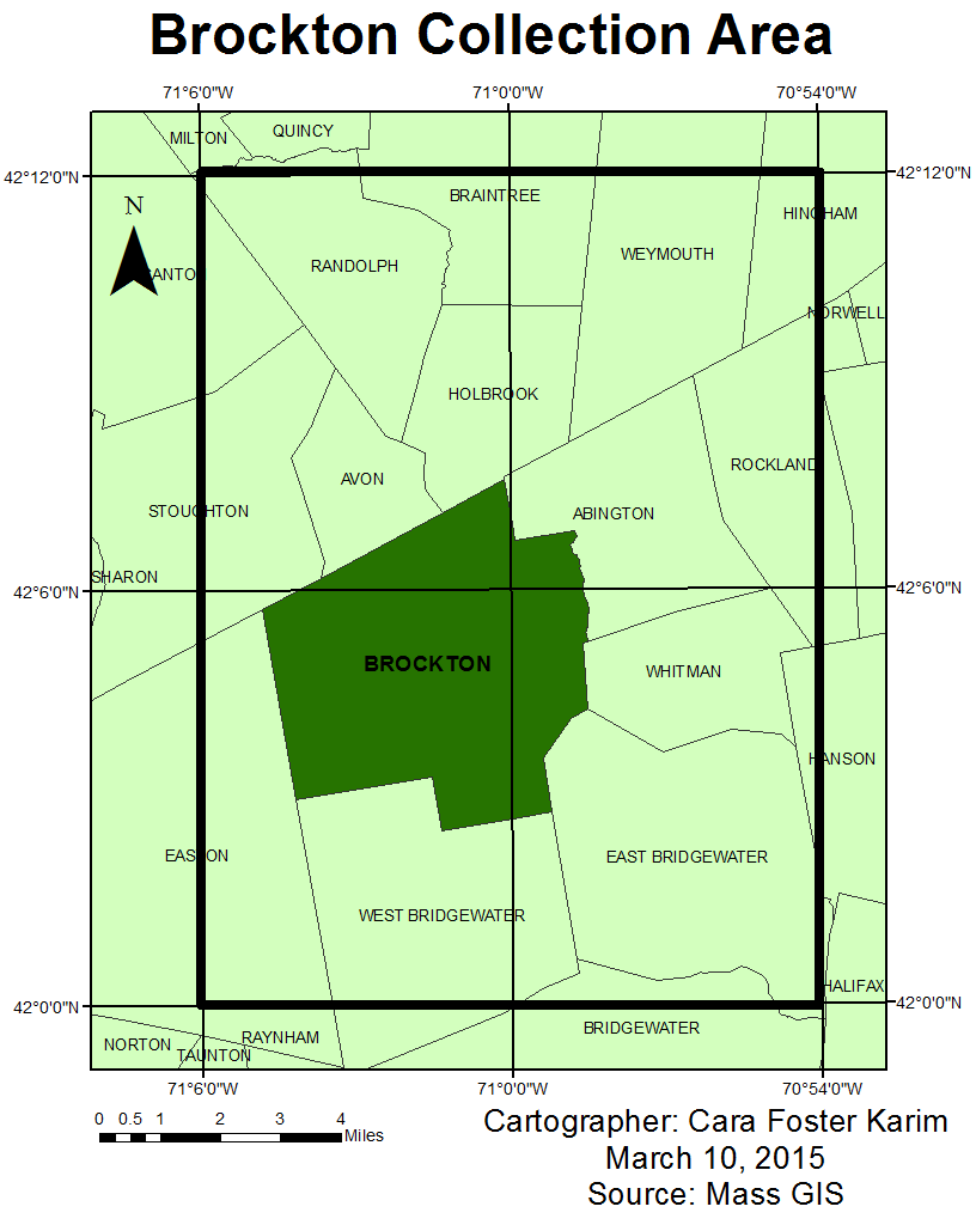
This area of research has great potential and there are a number of ways that future researchers could improve upon my methods in order to produce a study with fewer limitations. Comparing sentiment results between two different languages is inherently complicated due to different linguistic structures and cultural differences, but there are some steps that could be taken to mitigate these issues. First, using a computer program to identify the language of each Tweet and sorting the results into two separate

data sets before conducting any analysis would significantly eliminate the error associated with cognates in the sentiment lexicon resulting in Tweets being counted for Portuguese sentiment when not actually in Portuguese. Another solution for a researcher with a background in linguistics and psychology and a longer timeframe would be to edit or re-write both sentiment lexicons to ensure that they were as equivalent as possible in terms of numbers of words and the proportions of positive and negative words. This would also require more technical computer programming knowledge.

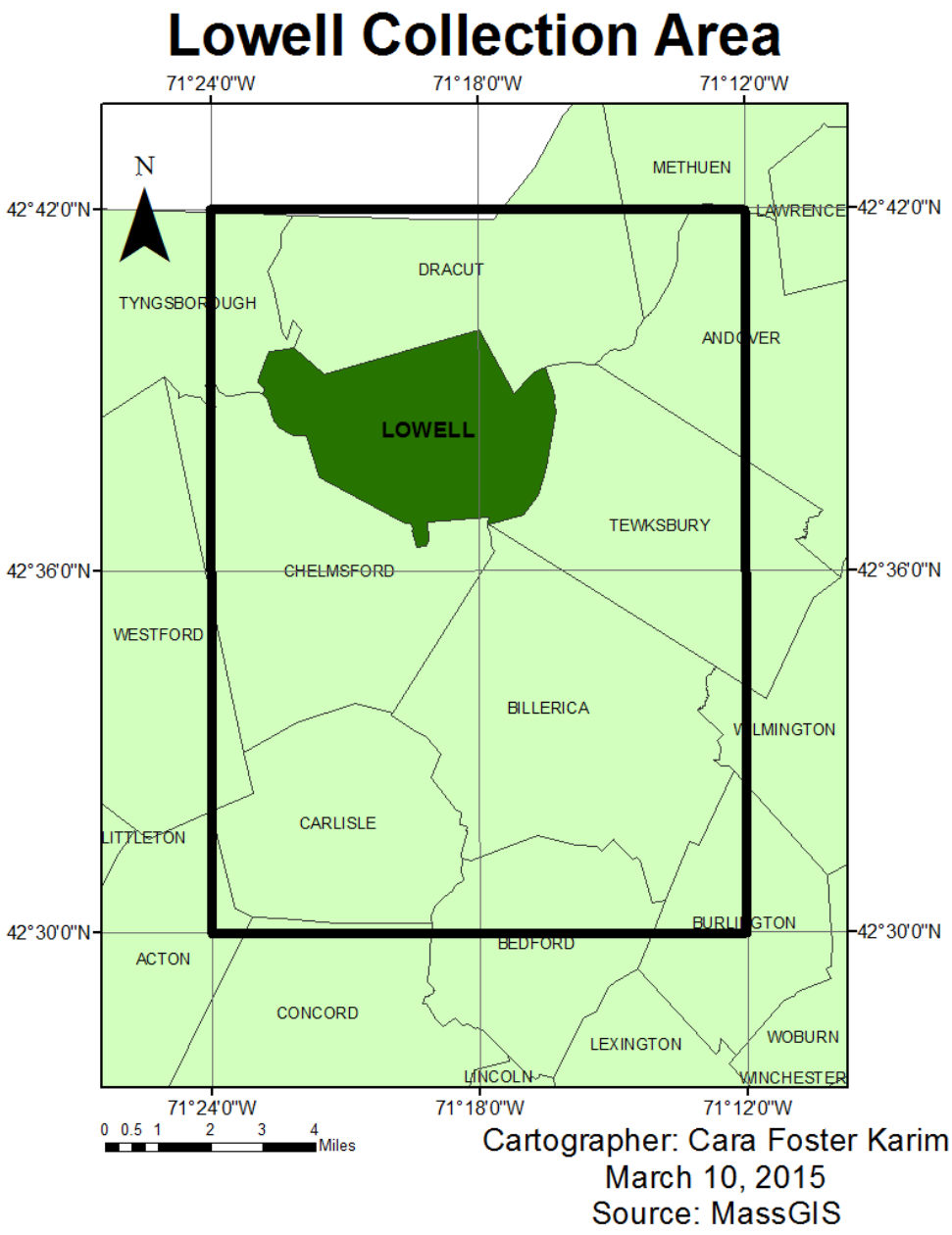
Another way to increase the relevance and validity of this study would be conduct a traditional survey of Portuguese and English speaking residents in these Gateway Cities to supplement the results from the sentiment analysis and the census data. Finally, increasing the sample size by including more cities and collecting data over a longer period of time would also increase the validity of the results. It would also be potentially beneficial to conduct a study in which Twitter sentiment results in different languages in a series of different international and domestic contexts would be compared, in an attempt to see how cultural differences or immigrant or native born status impact overall well-being. I look forward to seeing what direction future researchers take as they attempt to address similar research questions using sentiment analysis of microblog data.

Appendices

Appendix A: Map of Brockton data collection area



Appendix B: Map of Lowell data collection area



Appendix C: T-tests between proportions of positive Tweets in Portuguese results

	Eastern MA	**Brockton**	**Fall River**	**New Bedford**	**Lisbon**	**Sao Paulo**	**Harvard & MIT**	**Lowell**
Eastern MA								.
Brockton	Z-Score -0.1427, p-value 0.44433							
Fall River	Z-Score 0.0336, p-value 0.48803	Z-Score -0.0775, p-value 0.46812						
New Bedford	Z-Score 0.0002, p-value 0.5	Z-Score -0.047, p-value 0.48006	Z-Score -0.0249, p-value 0.49202					
Lisbon	Z-Score -4.5619, p-value 0	Z-Score 4.3704, p-value is 0	Z-Score 1.2699, p-value 0.10204	Z-Score 1.3175, p-value 0.09342				
Sao Paulo	Z-Score -7.8284, p-value 0	Z-Score 7.558, p-value 0	Z-Score -1.797, p-value 0.03593	Z-Score -1.8819, p-value 0.03005	Z-Score -3.0997, p-value 0.00097			
Harvard & MIT	Z-Score 1.4539, p-value 0.07353	Z-Score -1.4883, p-value 0.06811	Z-Score -1.1421, p-value 0.12714	Z-Score -1.1911, p-value 0.11702	Z-Score 2.4416, p-value 0.00734	Z-Score -2.866, p-value 0.00205		
Lowell	Z-Score -0.0109, p-value 0.49601	Z-Score -0.0447, p-value 0.48405	Z-Score -0.0249, p-value is 0.49202	Z-Score 0.0075, p-value 0.49601	Z-Score 1.562, p-value 0.05938	Z-Score -2.2493, p-value 0.01222	Z-Score -1.2664, p-value 0.10204	

Appendix D: T-tests between proportions of positive Tweets in English results

	Eastern MA	**Brockton**	**Fall River**	**New Bedford**	**Lisbon**	**Sao Paulo**	**Harvard & MIT**	**Lowell**
Eastern MA								
Brockton	Z-Score 0.3874, p-value 0.34827							
Fall River	Z-Score 4.6413, p-value 0	Z-Score 4.4253, p-value 0						
New Bedford	Z-Score 8.0452, p-value 0	Z-Score 7.7692, p-value 0	Z-Score 2.0976, p-value 0.01786					
Lisbon	Z-Score 12.2877, p-value 0	Z-Score 12.2002, p-value 0	Z-Score 9.8446, p-value 0	Z-Score 8.9547, p-value 0				
Sao Paulo	Z-Score 34.6994, p-value 0	Z-Score 33.656, p-value 0	Z-Score 15.3997, p-value 0	Z-Score 13.9761, p-value 0	Z-Score -3.8138, p-value 7E-05			
Harvard & MIT	Z-Score -9.0885, p-value 0	Z-Score -9.1635, p-value 0	Z-Score -10.4421, p-value 0	Z-Score -12.04, p-value 0	Z-Score -15.4439, p-value 0	Z-Score -20.5418, p-value 0		
Lowell	Z-Score 5.9444, p-value 0	Z-Score 5.6902, p-value 0	Z-Score 0.6054, p-value 0.27093	Z-Score -1.5762, p-value 0.05705	Z-Score -9.6709, p-value 0	Z-Score -15.8637, p-value 0	Z-Score 11.0646, p-value 0	

Appendix E: All results in Portuguese and English from all locations

Location	Language	Total # of tweets	Tweets w/ Sentiment	Coverage	Positive	Negative	% positive	% negative	Overall score	Score divided by # of Tweets with sentiment
Eastern MA	Portuguese	315448	4277	1.4%	2270	2199	50.8%	49.2%	-959	-0.2146
Fall River	Portuguese	30770	217	0.7%	112	109	50.7%	49.3%	-26	-0.1176
Brockton	Portuguese	273854	4211	1.5%	2235	2152	50.9%	49.0%	-958	-0.2184
New Bedford	Portuguese	37732	249	0.7%	128	124	50.8%	49.2%	-48	-0.1905
Lisbon	Portuguese	22156	7774	35.1%	4772	3907	55.0%	45.0%	-452	-0.0521
Sao Paulo	Portuguese	438048	154920	35.4%	99998	76451	56.7%	43.3%	26723	0.1514

Harvard & MIT	Portuguese	10505	134	1.3%	60	75	44.4%	55.6%	-44	-0.3259
Lowell	Portuguese	38663	350	0.9%	185	179	50.8%	49.2%	-89	-0.2445
Eastern MA	English	315448	169086	53.6%	113899	87523	56.5%	43.4%	61200	0.3038
Fall River	English	30770	17895	58.2%	11806	9700	54.9%	45.1%	1570	0.0730
Brockton	English	273854	144171	52.6%	96665	74471	56.5%	43.5%	48358	0.2826
New Bedford	English	37732	21783	57.7%	14262	12180	53.9%	46.1%	-1717	-0.0649
Lisbon	English	22156	2797	12.6%	1349	1629	45.3%	54.7%	1297	0.4355
Sao Paulo	English	438048	63966	14.6%	33031	34561	48.9%	51.1%	60408	0.8937

Harvard & MIT	English	10505	5533	52.7%	4026	2443	62.2%	37.8%	5504	0.8508
Lowell	English	38663	22037	57.0%	14454	12009	54.6%	45.4%	2561	0.0968

Appendix F: Words removed from Portuguese lexicon

blah	misses
bonus	moody
bug	nerd
bullshit	ok
cocksucker	okays
cool	oks
crap	omfg
crappy	omg
delicious	prod
douchebag	rock
fab	rofl
fucker	rude
fuckface	scariest
fucks	sexy
fuckwit	skank
gay	slur
geek	smart
h8	spam
haha	sucky
hehe	sup
lmao	super
lol	thanx
lolol	thx
luv	twat

ugh

wack

wow

wtf

x

xox

xx

yay

Appendix G: Correlation table

		Percent Positive English	Percent Positive in Portuguese	Percent Negative in English	Percent Negative in Portuguese	English score	Portuguese score	Normalized English score	Normalized Portuguese score
Percent Positive English	Correlation Coefficient	1	-0.1	-1.000**	0.3	-0.3	0.1	.900*	-0.5
	Sig. (2-tailed)	.	0.873	.	0.624	0.624	0.873	0.037	0.391
Percent Positive in Portuguese	Correlation Coefficient	-0.1	1	0.1	-.900*	.900*	-0.7	0	0.2
	Sig. (2-tailed)	0.873	.	0.873	0.037	0.037	0.188	1	0.747
Percent Negative in English	Correlation Coefficient	-1.000**	0.1	1	-0.3	0.3	-0.1	-.900*	0.5
	Sig. (2-tailed)	.	0.873	.	0.624	0.624	0.873	0.037	0.391
Percent Negative in Portuguese	Correlation Coefficient	0.3	-.900*	-0.3	1	. 1.000**	.900*	0.1	-0.1
	Sig. (2-tailed)	0.624	0.037	0.624	.	.	0.037	0.873	0.873
English score	Correlation Coefficient	-0.3	.900*	0.3	-1.000**	1	-.900*	-0.1	0.1
	Sig. (2-tailed)	0.624	0.037	0.624	.	.	0.037	0.873	0.873
Portuguese score	Correlation Coefficient	0.1	-0.7	-0.1	.900*	-.900*	1	-0.2	0.3
	Sig. (2-tailed)	0.873	0.188	0.873	0.037	0.037	.	0.747	0.624
Normalized English score	Correlation Coefficient	.900*	0	-.900*	0.1	-0.1	-0.2	1	-0.8
	Sig. (2-tailed)	0.037	1	0.037	0.873	0.873	0.747	.	0.104
Normalized Portuguese score	Correlation Coefficient	-0.5	0.2	0.5	-0.1	0.1	0.3	-0.8	1
	Sig. (2-tailed)	0.391	0.747	0.391	0.873	0.873	0.624	0.104	.
Total population	Correlation Coefficient	-0.1	-0.2	0.1	0	0	-0.3	0.3	-0.8
	Sig. (2-tailed)	0.873	0.747	0.873	1	1	0.624	0.624	0.104
Percent foreign born	Correlation Coefficient	0.5	-0.2	-0.5	0.1	-0.1	-0.3	0.8	-1.000**
	Sig. (2-tailed)	0.391	0.747	0.391	0.873	0.873	0.624	0.104	.

Median income	Correlation Coefficient	0.5	-0.2	-0.5	0.1	-0.1	-0.3	0.8	-1.000**
	Sig. (2-tailed)	0.391	0.747	0.391	0.873	0.873	0.624	0.104	.
Percent with high school diploma or higher	Correlation Coefficient	0.7	-0.1	-0.7	0	0	-0.4	.900*	-.900*
	Sig. (2-tailed)	0.188	0.873	0.188	1	1	0.505	0.037	0.037
Percent 18 and under	Correlation Coefficient	-0.3	.900*	0.3	-1.000**	1.000**	-.900*	-0.1	0.1
	Sig. (2-tailed)	0.624	0.037	0.624	.	.	0.037	0.873	0.873
Percent 65 and over	Correlation Coefficient	-0.5	0.2	0.5	-0.1	0.1	0.3	-0.8	1.000**
	Sig. (2-tailed)	0.391	0.747	0.391	0.873	0.873	0.624	0.104	.
Population density	Correlation Coefficient	0.2	-0.5	-0.2	0.3	-0.3	-0.1	0.5	-.900*
	Sig. (2-tailed)	0.747	0.391	0.747	0.624	0.624	0.873	0.391	0.037
Portuguese ancestry	Correlation Coefficient	-.900*	0	.900*	-0.1	0.1	0.2	-1.000**	0.8
	Sig. (2-tailed)	0.037	1	0.037	0.873	0.873	0.747	.	0.104
Brazilian ancestry	Correlation Coefficient	-0.3	.900*	0.3	-0.8	0.8	-0.6	-0.1	0.1
	Sig. (2-tailed)	0.624	0.037	0.624	0.104	0.104	0.285	0.873	0.873

**. Correlation is significant at the 0.05 level (2-tailed).*

***. Correlation is significant at the 0.01 level (2-tailed).*

Appendix H: Demographic variables by Portuguese and Brazilian ancestry for each city

Demographic variable	Fall River		New Bedford		Lowell		Brockton		Average		Average combined
	Brazilian	Portuguese	Brazilian	Portuguese	Brazilian	Portuguese	Brazilian	Portuguese	Brazilian	Portuguese	
% with high school diploma or higher	75.4%	58.5%	80.8%	59.3%	79.10%	67.10%	70.70%	72.40%	76.50%	64.33%	70.41%
Median household income	26,496	35,893	45,648	43,123	51,940	54,962	60,924	50,257	46,252	46,059	46,155
% under 18	18.9%	22.1%	29.2%	23.8%	29.02%	19.16%	26.09%	24.61%	25.79%	22.43%	24.11%
% 65 and over	4.3%	13.2%	3.0%	16.4%	0.00%	12.98%	0.00%	7.35%	1.47%	12.49%	6.98%
% Foreign born	78.9%	27.3%	72.7%	30.3%	76.1%	36.0%	73.1%	35.8%	75.2%	32.3%	53.78%

Sources: *United States Census Bureau (2015f, 2015g, 2015h, 2015i, 2015j)*

Appendix I: All Portuguese hand coding results

Tweet content	Sentiment words found by lexicon	Score	Sentiment words I identified	Adjusted score
Nao sei por que voce se foi"	Se (1)	1	não	-1
@cristywabi te odio	Odio (-4)	-4	odio	-4
O QUE TEM POR TRAS DESSE SORRISO QUE ME DEIXA FISSURADO EM VC @aissafranco_	Sorriso (2)	2	sorriso, fissurado	2
Grande grande #PedroMartinez http://t.co/NmMXQfX951	Grande (3)	3	grande	3
Sa se pawol Roro	Se (1)	1	se = neutral	0
@Adammes A Amanda. (Eu acho hahaha nao sei os nomes ainda direito)	Direito (-2)	-2	não, direito	-2
Vitim me chama visualiza dps de 15 dias ele responde	Chama (-3)	-3	dps (depois)--chama means "call"	-3
Obrigado Deus por mais um dia !#ObrigadoMeuDeus @ My House http://t.co/n3L3qGXfod	Obrigado (2)	2	obrigado	2
Friozinho bobo esse!!!! Silly cold!!! #Cold #Frio #BelowZero #Freezing #AbaixoDeZero @ Lowell ... http://t.co/KiggTUfoet	Bobo (-2), frio (-2)	-4	friozinho, bobo	-4
@luizanicolli eu sou do mal	Mal (-3)	-3	mal	-3
No filter !!!! #boston #ma #usa #winter #inverno #love #snow #neve #frio #Cold #gay #gaypic... http://t.co/VYARalKk5T	Frio (-2)	-2	"no filter" = usually positive	1
@_urgh Eu amava a parte do CHUTA ESSA PORRA HAHAHAHA	Porra (-1)	-1	amava, porra (profanity, but just used as intensifier here)	-1
My bebe sleeping	Bebe (2)	2	bebe	2
Ai a tatuagem do Caio Castro e temporaria. UFA!!!	Ai (-2)	-2	ai	-2
@felipeneto tava reclamando do frio no teu video olha hj como ta Kkk http://t.co/gTEAJq2csE	Frio (-2), como (2)	0	frio, reclamando	0
Espero muito que o Fernando mude a cabeca do Rafael. Ja que ele e o espertalhao. #BBB	Muito (3)	3	espero, muito	3
Dale Ucrania se sabia	Se (1)	1	se = if	0
Que os falsos vao tudo toma na porta do cu!	Falsos (-2)	-2	falsos	-2
Tio juninhoesj mordendo o pesinho da princesa e ela morrendo de rir pra ele! Ja estou com saudades da... http://t.co/7VIpM35uXI	Saudades (-3), rir (2)	-1	morrendo de rir = 4	2
Audio melhor http://t.co/jutsrsapoH	Melhor (2)	2	melhor	2
Tudo bem que era o video do Rodrigo Hilbert mas acho que assisti umas 3 vezes sem perceber	Bem (2)	2	tudo bem	2

@cristianesita ah mas tem gente que aguenta mais que a gente. Sempre tem. Ou que canaliza melhor as dores. Entende?	Dores (-4), melhor (2)	-2	dores, melhor	-2
Cara acho q nunca vi tanta neve na minha vida tipo nem em filme	Tipo (2)	2	neve	2
@luizanicolii dente dele igual uma placa de energia solar	Dente (-2)	-2	igual, energia	-2
Muito frio la fora e muitas bencao aqui dentro @ Revival Church for the Nations | Gloucester http://t.co/OFVfww8zdR	Frio (-2), muito (3)	1	bencao = 2	3
Ala a Tamires querendo champa e gritar UHU HAHAHAHA	Gritar (-4)	-4	querendo (gritar can mean to cry, but in this context just means shout)	1
Paseando por Boston !!! #snow#frio#examen#fisico#caminando#depinga#coffe Jumbo on track @ Scampo http://t.co/6n5RqCJlpP	Frio (-2)	-2	frio, !!	-2
Sorria mesmo que as circunstancias queira te fazer chorar @ Boston Massachusetts http://t.co/MfwrmoMexX	Chorar (-3)	-3	sorria, chorar	-3
Tao reprimindo a menina pq ela ta falando TETA. Imagina se ela falar ppk ne?	Se (1)	1	reprimindo = -3	-2
@jufurtado9 Ai achei que tava ficando louca miga HAHAHA ainda bem que alguem me entende.	Ai (-2), bem (2)	0	louca, entende, bem, ai	0
@VascainosF: Pode existir times grandes mas o Vasco e GIGANTE! /+/ http://t.co/A7jFkgKekl	Grandes (3)	3	grandes, gigante	3
Quem lhe deu essa sunga branca tava muito afim de lhe ver de sunga branca Mariza sendo gente como a gente falando com Fernando #BBB	Muito (3), como (2)	5	muito, afim	5
Manel melhor manel kkkkkkkkkkkkkkk	Melhor (2)	2	melhor	2
O melhor lugar do mundo e dentro de um abraco	Melhor (2)	2	melhor, abraco	2
@gui_sad to fragano c falano mal d mim no wpp pro gabriel ta cabo nosso laco amizadezal	Mal (-3)	-3	mal	-3
Como as criancas sao belas ce tenta fazer um video fofo de um falando seu nome e o outro ta fazendo forca pra fazer coco atras. HAHAHAHA	Como (2)	2	belas, criancas, fofo, coco	2
WhatsApp vai tomar no cu querido. Obrigado de nada.	Querido (2), obrigado (2)	4	obrigado de nada = -4	-2
@Gi_cercadinho foi muito lindo doi ne	Muito (3)	3	muito, lindo, doi	3
Nada e errado se te faz feliz	Feliz (2), errado (-2)	0	feliz	0
Link e Bowser entram na batalha! #amiibo http://t.co/OhjTReKgmJ	Batalha (-2)	-2	Link, Bowser = 4	2
nada e capaz de apagar as lembrancas que voce deixou	Apagar (-2)	-2	nada, capaz, lembrancas	-2
Ainda sem entender o que o Fernando fez excluindo as gurias. serio. Babaca. #BBB	Babaca (-3)	-3	excluindo, babaca	-3

@KatheTorresR se Arabe dificil	Se (1)	1	dificil = -3	-2
Voces nao tem nocao de como minha vida mudou	Como (2)	2	não, vida,	2
@choracuica @goiabatobias Eu certamente ia morrer sem saber hahahaha	Morrer (-2)	-2	morrer, haha	-2
@leorodriguezbr boa noite Leozin que papai do ceu te abencoe sempre	Boa (2)	2	boa, abencoe	3
Turistando pela cidade! @ Ha'penny Bridge - Dublin http://t.co/1RLOdOjmSN	Ha (2)	2	turistando	2
Bom Diahoje apareceu um solzinho para esquentar mas tem muita neve para limpar... http://t.co/XbKZxt0TPf	Bom (2)	2	bom, neve	2
@lucasvernek @cajuetotonho Eu amanheci aqui coladinho em vcs mais veja a paisagem la fora temperatura -7 ainda bem http://t.co/drqmwrL2dK	Bem (2)	2	bem	2
@DParticular acho que nao hahahaha	Nao (-3)	-3	não, haha	-3
ESSE PORRA DESSE ADRILAHEUSB NAO DEIXA OS OUTROS FALAREM PLMDDS TIRA ELE.	Porra (-1)	-1	porra, nao	-1
@MLSBrasil ta fazendo Falta aqui Heim Thiago.	Falta (-2)	-2	falta	-2
Errar e o mano	Errar (-2), mano (2)	0	errar	0
Quero ser feliz tambem navegar nas aguas do teu marrrr	Feliz (2)	2	feliz	2
Pior que o chato nunca sai ne gente. Fico besta que a gente tem que ficar aguentando.	Pior (-3), chato (-2), besta (-2)	-9	pior, chato	-9
E meio estranho saber que o povo aqui nao me conhece de verdade. Eles nao sabem que eu escuto musica ruim por exemplo.	Verdade (2), estranho (-2), ruim (-2)	0	conhece, verdade, ruim	0
Descanse em paz Marcinho Meus sentimentos a familia sua. Voce era um pai excelente tio e pessoa. Foi uma honra conhece-lo	Paz (1), honra (1)	2	paz, honra, excelente	2
Private nao Se va p ara fora http://t.co/Tr4ufyAPUo	Nao (-3), se (1)	-2	não	-2

Appendix J: English hand coding results

Please note that for the sake of brevity, only the results for which I had to adjust the score are included here. The ones that I assessed to be correct already are not displayed.

Tweet content	Sentiment words found by lexicon	Score	Sentiment words I identified	Adjusted score
Not sure the gate agent chose the best words; There's a hole in the airplane they are trying to close.	best(3)	3	best, hole, not	-2
@CAStrunk @krista_0x @BGiarla guys i'm only like 10 stop swearing thx	swearing (-2), stop (-1), like (2)	-1	like in filler sense = 0	-3
@MattSBrodsky @mikefreemanNFL Must be cheating. Check all Pats offensive players hands for web like fingers.	like(2)	2	cheating = -3, like in neutral sense = 0	-3
Hate going home late cause of my coworkers. Like hurry up.	hate (-3), like (2)	-1	like in filler sense = 0	-3
@_bdsxo @Pointdexter I saw that lmao	lmao (4)	4	lmao = 1	1
Don't like this Wendys or the people in it	like (2)	2	"don't like" = -2	-2
The feeling of not being good enough is eating me alive and destroying me legit	destroying (-3), good (3), feeling (1), alive (1)	2	"not...good" = -2, "eating me alive" = -3	-8
Things that should stress me out: Grades. Internships. Money. Things that do stress me out: Likes/ Favorites on social media #priorities	likes (2), favorites (2)	4	stress = -3	1
Nobody wants to play GTA with me ever	wants(1)	1	"nobody" = -2	-1
Nothing like having to pull out 2 full skeins of work when you realize you dropped a stitch (or 3) at.. http://t.co/jomm43yrXy	like (2)	2	like in neutral sense = 0	0
Hold your finger in a tweet and it puts it immediately in a dm for someone. Its like twitters beggin ppl to @ these hoes on their tweets	like (2)	2	like in neutral sense = 0	0
@velasquezDO You can't guard me d! Lol u gotta give me space or ima eat u to the hoop and if u give me space I'm splashing u seen it today	lol (3)	3	lol = 1	1
I honestly can't play with people I don't know I'm done giving kids i don't know chances like If I don't know you don't ask to play with me	chances (2), like (2)	4	can't, done, don't = -3	-3
@caamxo: It pisses me off when people say y'all lol idk why sammme	lol (3)	3	lol = 1	1
Finals got me like http://t.co/KDIDCwe6Tk	like(2)	2	like in neutral sense = 0	0
@CreepyDemonLord // Ghira promise me you won't hurt him again...// *smiles*	smiles (2), won (3), promise (1), hurt (-2)	4	"won't" =/ "won"	1
@theblaze Never going to happen so who cares? Unfortunately.	cares (2)	2	unfortunately = -2, never = -3	-3

@GenyFrumDaBlock bless ur soul lmao	bless (2), lmao (4)	6	lmao = 1	3
@Fvck_AJ @Diamonx3 @MirandaaO @Gousby_617 nobody cares	cares (2)	2	nobody = -3	-1
@OmvrRivera: : You have such a clear face it's adorable asf but you never hmu thank you lol	clear(1), adorable (3), thank (2), lol (3)	9	lol = 1	7
#RIPCaroline heaven gained a beautiful angel!! You're gonna be missed by so many!! Fly High beautiful	gained (2), beautiful (3), heaven (2), missed (-2)	5	RIP = -4	1
@g0dsbitch you win lmao	win(4), lmao (4)	8	lmao = 1	5
I wanna go back to Old me and I won't tell nobody	won (3)	3	"won't" =/ "won"	0
What's it like to stare into the eye of Satan's butthole	like (2)	2	satan= -3, butt = -2	-5
And I also ordered riff raff tickets lmao	lmao (4)	4	lmao = 1	1
@nicklicciardi12 got that right we gonna kill it doe.. love you a&f	love (3), kill (-3)	0	kill in positive sense =2	5
I have a massive crush on peter capaldi sorry nawt sorry	crush (-1), sorry (-1)	-2	crush in positive sense = 2	1
And you thought #riskmanagement jobs were hot. Hello #cybersecurity RT @kevinroose: Latest @SonyPictures data dump http://t.co/50PP75nhgn	dump (-1)	-1	hot = 3	2
Cannot wait for 50 shades of grey to come out	grey (-1)	-1	"cannot wait" = 3	2
Some 80 year old lady just fist bumped me. How the fuck is your day going?	fuck (-4)	-4	fist bump = 3?	-1
my work is gonna pay for me to take a baking class!!	pay (-1)	-1	baking = 3	2
@ryan_palmer5: This is my bestfriend http://t.co/mtZw5SdVPv omg I love this//you	love (3)	3	bestfriend, love	5
@SHEEPISHLORDOFC LMFAOOO that baby is mad cute	cute (2), mad (-3)	-1	mad as intensifier = 1	3
Today would be a great day to find $20 on the ground	great (3)	3	great	3

Bibliography

Antonelli, F., Azzi, M., Balduini, M., Ciuccarelli, P., Valle, E. D., & Larcher, R. (2014, May). City sensing: visualising mobile and social data about a city scale event. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (pp. 337-338). ACM.

Araújo, M., Gonçalves, P., & Benevenuto, F. (2013) Métodos para análise de sentimentos no Twitter. Presented at *WebMedia'13*, November 5–8, 2013, Salvador, Brazil.

Bailey, B. (2000). History and description of Portuguese immigration and the East Providence/SE New England Portuguese community. University of Massachusetts – Amherst. Retrieved from the Internet March 31st, 2015 at: http://works.bepress.com/benjamin_bailey/63

Balage Filho, P. P., Pardo, T. A., & Aluısio, S. M. (2013) An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, (pp.215–219) Fortaleza, CE, Brazil, October 21–23, 2013. c 2013 Sociedade Brasileira de Computação

Balduini, M., Della Valle, E., Dell'Aglio, D., Tsytsarau, M., Palpanas, T., & Confalonieri, C. (2013). Social listening of city scale events using the streaming linked data framework. In *The Semantic Web–ISWC 2013* (pp. 1-16). Springer Berlin Heidelberg.

Ben-Harush, O., Carroll, J. A., & Marsh, B. (2012). Using mobile social media and GIS in health and place research. *Continuum*, *26*(5), 715-730.

Bertrand, K. Z., Bialik, M., Virdee, K., Gros, A., & Bar-Yam, Y. (2013). Sentiment in New York City: A high resolution spatial and temporal view. New England Complex Systems Institute, Cambridge, MA.

Bollen, J., Pepe, A., & Mao, H. (2011, July). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. 5th International AAAI Conference on Weblogs and Social Media [ICWSM].

Clayton-Matthews, A. and Watanabe, P. (2012). Massachusetts immigrants by the numbers, Second Edition: Demographic characteristics and economic footprint. Immigrant Learning Center, Inc. Retrieved from the internet December 1st, 2014 at:http://www.ilctr.org/wp-content/uploads/2013/05/ILC-Mass-Immigrants-by-the-Numbers-2nd-ed.pdf

Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual Twitter sentiment analysis. In *Information retrieval technology* (pp. 238-249). Springer Berlin Heidelberg.

de Araujo, G. D., Sousa, F. S., Teixeira, F., Mancini, F., De Domenico, E. B. L., de Paiva Guimarães, M., & Pisa, I. T. (2012). Análise de sentimentos sobre temas de saúde em mídia social. *Journal of Health Informatics*, *4*(3).

Diener, E., & Seligman, M. E. (2004). Beyond money toward an economy of well-being. *Psychological science in the public interest*, *5*(1), 1-31.

Feldman-Bianco, B. (1992). Multiple layers of time and space: The construction of class, ethnicity, and nationalism among Portuguese immigrants. *Annals of the New York Academy of Sciences*, *645*(1), 145-174.

Foreman, B. and Larson, S. (2014). Going for growth: Promoting immigrant entrepreneurship in Massachusetts Gateway Cities. MassINC Gateway Cities Innovation Institute. Retrieved from the internet December 1st, 2014 at: http://www.massinc.org/~/media/Files/Mass%20Inc/Research/Full%20Report%20PDF%20files/GoingforGrowth_PromotingImmigrantEntrepreneurshipinMassachusettsGatewayCities.ashx

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2013). Sensing urban land use with Twitter activity. Telefonica Research, Madrid, Spain.

Fujisaka, T., Lee, R., & Sumiya, K. (2010, January). Discovery of user behavior patterns from geo-tagged micro-blogs. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication* (p. 36). ACM.

Gao, Q., Abel, F., Houben, G. J., & Yu, Y. (2012). A comparative study of users' microblogging behavior on Sina Weibo and Twitter. In *User modeling, adaptation, and personalization* (pp. 88-101). Springer Berlin Heidelberg.

Garas, A., Garcia, D., Skowron, M., & Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports*, 2, article 402.

Giannopoulos, G., Weber,I., Jaimes,A., Sellis,T. (2012).Diversifying user comments on news articles, web information systems engineering (WISE 2012). Lecture Notes in Computer Science 2012, pp 100-113.

Gonçalves, P., Dores, W., & Benevenuto, F. (2012). PANAS-t: Uma escala psicométrica para medição de sentimentos no Twitter. Departamento de Computação, Universidade Federal de Ouro Preto (UFOP).Ouro Preto, MG, Brasil.

Gordon, J. (2013). Comparative geospatial analysis of Twitter sentiment data during the 2008 and 2012 US Presidential elections. University of Oregon.

Grigore, M. and Rosenkranz, C. (2011). Increasing the willingness to collaborate online: an analysis of sentiment-driven interactions in peer content production. ICIS 2011 Proceedings. Paper 20.

Hale, S. A. (2014, April). Global connectivity and multilinguals in the Twitter network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 833-842). ACM.

Hollander, J., Graves, E., Levanthal, T. (2014). Using big data to study urban sentiments: Twitter data vs. published meeting minutes. Tufts University.

Immigrant Learning Center. (2010). Massachusetts immigrant entrepreneurs: Engines for economic growth, wealth and job creation. The Immigrant Learning Center, Inc. Retrieved from the internet December 1st, 2014 at:http://www.ilctr.org/wp-content/uploads/2010/11/ILC-Profiles-web-2.pdf

Jouët-Pastré, C., & Braga, L. J. (2005). Community-based learning: A window into the Portuguese-speaking communities of New England. *Hispania*, 863-872.

Kramer, A. D. (2010, April). An unobtrusive behavioral model of gross national happiness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 287-290). ACM.

Kucuktunc, O., Cambazoglu, B.B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! Answers, Proceedings of the 5th ACM International Conference on Web Search and Data Mining.

Leavy-Sperounis, M. M. B. (2010). *Manufacturing recovery: a networked approach to green job creation in Massachusetts Gateway cities* (Doctoral dissertation, Massachusetts Institute of Technology).

Lima, A., & Siqueira, C. E. (2007). Brazilians in the US and Massachusetts: a demographic and economic profile. Gastón Institute Publications. Paper 50.
Retrieved from the internet March 31st, 2015 at:
http://scholarworks.umb.edu/gaston_pubs/50

Lovelace, R., Malleson, N., Harland, K., & Birkin, M. (2014). Geotagged Tweets to inform a spatial interaction model: A case study of museums. School of Geography, University of Leeds, Leeds, UK.

MacEachren, A. M., Robinson, A. C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J., & Mitra, P. (2011, July). Geo-twitter analytics: Applications in crisis management. In *25th International Cartographic Conference* (pp. 3-8).

Magdy, A., Ghanem, T. M., Musleh, M., & Mokbel, M. F. (2014). Exploiting geo-tagged Tweets to understand localized language diversity. In Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data(p. 2). ACM.

Massa, P., & Campagna, M. (2014). Social media geographic information: Recent findings and opportunities for smart spatial planning. *Tema. Journal of Land Use, Mobility and Environment*.

Massachusetts Alliance of Portuguese Speakers. (2008). In Massachusetts. Retrieved from the internet November 6th 2014 at: http://www.maps-inc.org/our-community/portuguese-speaking-community-in-massachusetts/

Massachusetts Alliance of Portuguese Speakers. (2012). Community health profile . Retrieved from the internet November 6th 2014 at: http://www.maps-inc.org/wp-contenuploads/2012/05/Community-Health-Profile.pdf

McGowan, O. T. (1976). Factors contributing to school leaving among immigrant children: The case of the Portuguese in Fall River, Massachusetts. PhD dissertation. Catholic University of America, Washington, D.C..

Mearns, G., Simmonds, R., Richardson, R., Turner, M., Watson, P., & Missier, P. (2014). Tweet my street: A cross-disciplinary collaboration for the analysis of local Twitter data. *Future Internet*, *6*(2), 378-396.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *ICWSM*, *11*, 5th.

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE 8(5): e64417. doi:10.1371/journal.pone.0064417*

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: Mapping world languages through microblogging platforms. *PloS one*, *8*(4), e61981.

Nishikawa, A. (2014, September 9). Twitter streaming by locations [Msg 1]. Message posted to http://stackoverflow.com/questions/23318112/twitter-streaming-by-locations

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, *11*, 122-129.

Pap, L. (1981). *The Portuguese-Americans*. Twayne Pub

Pfitzner, R., Garas, A., & Schweitzer, F. (2012). Emotional divergence influences information spreading in Twitter, ICWSM-12.

Poblete, B., Garcia, R., Mendoza, M., & Jaimes, A. (2011, October). Do all birds Tweet the same?: Characterizing Twitter around the world. In

Proceedings of the 20th ACM International Conference on Information and Knowledge Management(pp. 1025-1030). ACM.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012, February). Tracking gross community happiness from Tweets. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 965-968). ACM.

Santos, L. M., Esmin, A. A. A., Zambalde, A. L., & Nobre, F. M. (2011). Twitter, análise de sentimento e desenvolvimento de produtos: Quanto os usuários estão expressando suas opiniões?. *Revista PRISMA. COM*, (13).

SentiStrength. (2015). About SentiStrength. Retrieved from the internet March 31st, 2015 at: http://sentistrength.wlv.ac.uk/#About

Silvia, P. T. (1976). The position of "new" immigrants in the Fall River textile industry. *International Migration Review*, 221-232.

Siqueira, C. E., & Jansen, T. (2012). Working conditions of Brazilian immigrants in Massachusetts. *Journal of Immigrant and Minority Health*, *14*(3), 481-488.

Smith, A., & Brenner, J. (2012). Twitter use 2012. *Pew Internet & American Life Project*, 4.

Souza, M., & Vieira, R. (2012). Sentiment analysis on Twitter data for Portuguese language. In *Computational Processing of the Portuguese Language* (pp. 241-247). Springer Berlin Heidelberg.

Souza, M., Vieira, R., Busetti, D., Chishman, R., & Alves, I. M. (2011). Construction of a Portuguese opinion lexicon from multiple resources. *STIL*.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology,* 62(2), 406-418.

Twitter, Inc. (2015). Streaming API request parameters: Locations. Retrieved from the internet March 11th, 2015 at: https://dev.twitter.com/streaming/overview/request-parameters#locations

UMassDartmouth. (2008).The Gateway Cities compact for community and economic development. Retrieved from the internet November 6th 2014 at: http://www.umassd.edu/media/umassdartmouth/seppce/centerforpolicyanalysis/urbaninitiative/gatewaycities/signedcompact.pdf

UMass Dartmouth. (2014a)The Urban Initiative at UMass Dartmouth. Retrieved from the internet November 6th 2014

at:http://www.umassd.edu/media/umassdartmouth/seppce/centerforpolicyanalysis/urbaninitiative/assets/urbaninitiative.pdf

UMass Dartmouth. (2014b). Plan for renewal. Retrieved from the internet November 6th 2014 at: http://www.umassd.edu/media/umassdartmouth/seppce/centerforpolicyanalysis/urbaninitiative/assets/planforrenewal.pdf

United States Census Bureau.(2014a). Fall River (city) State and County QuickFacts. Retrieved from the internet December 3rd, 2014 at: http://quickfacts.census.gov/qfd/states/25/2523000.html

United States Census Bureau.(2014b).New Bedford (city) State and County QuickFacts. Retrieved from the internet November 6th 2014 at:http://quickfacts.census.gov/qfd/states/25/2545000.html

United States Census Bureau. (2014c). Selected Characteristics of the Native and Foreign Born Populations, 2008--2012 American Community Survey 5-year estimate.

United States Census Bureau.(2015a). Cambridge (city) State and County QuickFacts. Retrieved from the internet March 11th 2015 at http://quickfacts.census.gov/qfd/states/25/2511000.html

United States Census Bureau.(2015b). Lowell (city) State and County QuickFacts. Retrieved from the internet March 11th 2015 at http://quickfacts.census.gov/qfd/states/25/2537000.html

United States Census Bureau.(2015c). Brockton (city) State and County QuickFacts. Retrieved from the internet March 11th 2015 at http://quickfacts.census.gov/qfd/states/25/2509000.html

United States Census Bureau. (2015d). Table B01003. 2006-2010 American Community Survey Selected Population Tables.

United States Census Bureau. (2015e). 2009-2013 American Community Survey Selected Population Tables.

United States Census Bureau. (2015f). DP02 selected social characteristics in the United States. 2006-2010 American Community Survey Selected Population Tables.

United States Census Bureau. (2015g). B01001 Sex by age. 2006-2010 American Community Survey Selected Population Tables.

United States Census Bureau. (2015h). B19013 Median household income in the past 12 months (in 2010 inflation-adjusted dollars). 2006-2010 American Community Survey Selected Population Tables.

United States Census Bureau. (2015i). CP02 Comparative social characteristic in the United States. 2013 American Community Survey 1-year Estimates.

United States Census Bureau. (2015j). B05002 Place of birth by citizenship status. 2006-2010 American Community Survey Selected Population Tables.

Van Kamp, I., Leidelmeijer, K., Marsman, G., & De Hollander, A. (2003). Urban environmental quality and human well-being: Towards a conceptual framework and demarcation of concepts; a literature study. *Landscape and urban planning*,*65*(1), 5-18.

Vural, G., Cambazoglu,B. B., and P. Senkul (2012). Sentiment-focused web crawling, Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp 2020-2024.

Weber, I., Ukkonen, A., & Gionis, A. (2012). Answers, not links: extracting tips from Yahoo! answers to address how-to web queries. Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12).

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. *Journal of the American Society for Information Science and Technology*, *63*(8), 1631-1646.

Williams, J. R. (1982). *And yet they come: Portuguese immigration from the Azores to the United States.* Center for Migration Studies, 209 Flagg Place, Staten Island, NY 10304.

Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the "geographical awareness" of individuals: an exploratory analysis of Twitter data. *Cartography and Geographic Information Science*, *40*(2), 103-115.

Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets?. *Scientific Reports*, *4*. doi:10.1038/srep04213.