Understanding the Population of a Modern City Through Twitter

Josefin Morelius[1,1*] and Stephanie Cao[1,1†]

*Corresponding author(s). E-mail(s): moreliusjosefin@gmail.com;
Contributing authors: caosteph@seas.upenn.edu;
†These authors contributed equally to this work.

Abstract

This paper examines the use of Twitter data to understand the interests of a modern Spanish city. The study focuses on extracting the most popular topics discussed by the citizens in Valencia by applying the natural language processing technique, topic modeling, to process and classify the tweets. The paper also includes an analysis of the most frequent words and topics discussed on Twitter by the users in the city, and what insights they may provide to city decision-makers. The study demonstrates the potential of social media data analysis in understanding urban populations and informing policy-making decisions. The results indicate that the frequency of some of the topics that are being discussed on Twitter reflect the ongoing events in Valencia, however, the majority remain stable over time. The findings of this research have implications for urban planners, marketers, and policymakers in developing effective strategies for engaging with city residents. In this paper sentiment analysis is not being discussed which could be an area for further research.

Keywords: Valencia, Twitter, Social Media, Decision making, Topic Modeling, Smart City

1 Introduction

A fast-forward digital transformation has been made during the last few decades. As a consequence, the number of Smart cities has increased tremendously. The use of technology within smart cities has been seen as one of the main pillars of the smart city concept in order to address some problems

in urban environments, such as emergency response, energy efficiency, public security, and mobility. Meanwhile, our society is becoming more and more digital, and the number of people using social media is constantly increasing. Different social platforms are used for unique purposes, but Twitter is preferable since the main purpose is to share and seek information, it is not to interact with friends in the same way as other social media platforms. Therefore, it is clear that social media posts, particularly from Twitter, can be an effective tool for smart city decision-makers to learn more about their citizens [1]. Additionally, this source is a free, public, and easily accessible way in order to conduct research on the public, allowing the government to gain insights into public sentiment and adjust its policies and messaging accordingly. It can be beneficial when trying to assess citizen concerns in real-time when making time-sensitive decisions. Furthermore, social media posts are generally more unbiased and unfiltered than information gathered through other sources.

In the past decade, much research has focused on using Twitter as a tool for event detection in order to optimize smart city networks [7]. Previous research has also discussed use cases like more efficient disaster response and incident reporting [12]. However, it still remains unclear how smart cities can analyze tweets to better understand the concerns, opinions, and feelings of their citizens regarding different policies, allowing them to make better-informed decisions.

The purpose of this study is to allow decision-makers to take better decisions proactively or at early stages by leveraging Twitter data. Additionally, we will gain a better understanding of public opinion since social media data can help the government to understand what people are saying about specific policies, political issues, and government officials and adjust its policies and messaging accordingly. The government can also use social media data to evaluate public response to policy initiatives and services, which can help make data-driven decisions on where to allocate resources and improve service delivery.

1.1 Aims

In this study, we will employ the natural language processing (NLP) technique of topic modeling in order to analyze a large data set of Tweets including the word or hashtag "Valencia." Topic modeling will allow us to extract the most relevant topics being discussed by Valencian citizens. By applying this analysis to the entire data set in addition to the most recently posted Tweets, we can determine what issues have been around for a long time, and which have arisen only recently and might be more time sensitive. Thus, we will be able to gain insight into long-term and short-term issues, as well as learn about what citizens are already satisfied with.

1.2 Results

From our analysis, we learn about what major public issues citizens are talking the most about, which include public health, public transportation and traffic,

and corruption within local politics. Additionally, we gain insight into other ways that Valencia citizens use Twitter, which include expressing thanks and gratitude, as well as sharing pictures of their lives.

1.3 The Structure

First, this paper will conduct a complete review of the existing work regarding the application of social media in smart cities and smart city decision-making processes. In this section, we will also define all relevant terms and concepts. Next, we will detail our methodology for obtaining the data used, including the rationale behind the data-cleaning process. Then, we will explain how we employ topic modeling to find the most popular subjects discussed. Finally, we display and analyze the results obtained in this study and finish with our concluding remarks.

2 Background

In this section, theory and related research on how a smart city can use social media, more specifically Twitter, in order to understand the citizens better will be presented.

2.1 Definitions

2.1.1 Smart city concept

The "smart city" term lacks a clear definition, which complicates the process of becoming a smart city.[9] However, in this paper, the definition of a smart city will be referred to as a city with technology-based infrastructure, which enables it to increase operating efficiency and address social, economic, and environmental issues on different levels.

2.1.2 Decision

While decisions can be defined in a multitude of ways, in this study we will define it as the process of choosing a course of action based on a set of criteria and objectives. It can also be viewed as some "action" or "change" that is made. In the context of smart cities using social media to make decisions, a decision could be the selection of a particular policy, strategy, or action based on the analysis of social media data.

2.2 Twitter Data set

Twitter has over 400 million monthly active users who collectively post 500 million tweets (posts on Twitter) daily. [11] The tweets consider different everyday activities making it a very attractive database to use when studying people's behavior in different viewpoints and contexts. The opportunity to get real-time data is another parameter that makes the use of a Twitter data set even more interesting. The use of real-time data might be useful to understand citizen

concerns in real-time when trying to make time-sensitive decisions. Therefore, the use of Twitter as an early warning system for emerging issues and trends can be very useful. In one article, a system detected nearly all the earthquakes by just monitoring tweets [3].

Twitter is also considered to be more informational compared to a lot of other social media platforms since people use it more as a tool to spread and gain information. The result of this use makes the output of the tweets more negative than other platforms since people post on Twitter to express themselves rather than show off the best parts of their life. Users post more quickly compared to Instagram and TikTok where people might wait a bit to post for the right moment.

2.3 Topic Modeling

Topic modeling is a technique that can help determine the main topics of a body of text. By using this type of algorithm it is possible to identify underlying themes or topics in a corpus of documents that may not be immediately apparent to human readers. This can provide new perspectives and insights into the content of the text. Topic modeling has been successfully applied in a wide range of domains, including social media analysis, healthcare, and finance. By leveraging domain-specific knowledge, topic modeling can provide valuable insights into text data that may not be possible through other methods. Additionally, topic modeling can generate visualizations that help to interpret and understand the relationships between topics and documents. These visualizations can aid in the communication of the findings to a broader audience [4].

2.4 Searching for the Best References

To conduct the literature review, Google Scholar was used as the primary search engine. In order for it to be comprehensive, a variety of keywords and combinations thereof were used, including, but not limited to, “smart city,” “social media,” “Twitter,” and “decision-making.” The articles found were rated from 1 (least relevant to our topic) to 5 (most relevant), considering factors such as the fact that the data set used in this study did not contain any geographical information per tweet.

2.5 Analyzing the References

After reading the relevant works using the previously described method, we came across several ways that the analysis of social media, as well as specifically with tweets, can be applied to smart cities or city governance.

We found that event detection was a common application of tweet analysis that has been studied in this field. Event detection can be very useful for cities to help with controlling crowds, optimizing their sensor networks, and implementing traffic-reduction measures. However, one main constraint was that our data did not contain any geographical data from the tweets that are

posted. While this poses a substantial limitation on the scope of our analysis, since it cannot be used for useful applications like event detection, it also means that our results can be obtained with devices that have small storage limits and thus is much more accessibly computed.

Thus, we focused on articles that explored social media analysis applications that did not utilize location-based analysis. Sentiment analysis was continually mentioned as a useful tool specifically for tweets, as it is the most popular platform for people to express opinions in their own words. The use of social media to ascertain public opinion and citizen happiness was also brought up in the studies that were reviewed, reaffirming our hypothesis that Twitter data will be a critical tool for smart city decision-makers to utilize.

2.6 Synthesis

While many existing works utilize location-based data from social media posts, The New Eye of Smart City: Novel Citizen Sentiment Analysis in Twitter [8] does discuss the use of sentiment analysis for mood monitoring in smart cities, using data from two months of tweets posted from New York City. The study was able to monitor changes in citizens' moods throughout the course of the day, and also offer some useful guidelines for our methodology. We hope to expand on their work by examining data from over a longer period of time, as well as examining day-to-day changes in mood. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis[7] not only applies sentiment analysis but also an LDA model for topic modeling in order to determine what subjects citizens are discussing the most. From these studies, we can learn the data processing and analysis techniques that can form our starting point.

3 Framework

3.1 Elements

The data required for our project is the set of Tweets with the hashtag #Valencia. It will be provided. The data provider would be the Twitter application, where the Tweets will be collected from. The users will also be used as sensors since they will provide the information. Our system needs users to be able to post Tweets.

3.2 Processing

First, we want to store all the information in a CSV file containing the user, time, text, likes, retweets, and replies for each Tweet. Then, we will clean this dataset, for instance, for duplicates, URLs, get all of them in the same language, and pass them all to lowercase, in order to extract unnecessary information and fake news. We will then remove the user attribute from each Tweet as that only represents an id.

3.3 Conceptual Model

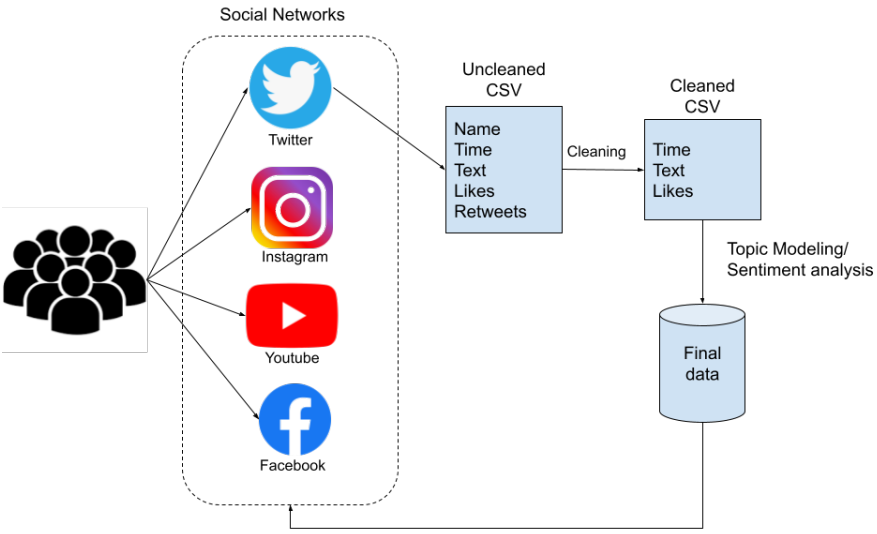

Fig. 1 The Conceptual Model of the process.

4 Method

4.1 Data needs

We want to help decision-makers in Valencia optimize their process for solving problems and assessing citizens' opinions on public issues. We also want to allow decision-makers to gauge the mood and activity of citizens in real time.

We need a dataset of the time, text, and likes of tweets that pertain to Valencia (have the hashtag #Valencia).

We have a dataset of tweets with the hashtag #Valencia, with the following attributes: name, time, text, likes, and retweets. We can clean this data and extract the important columns in order to obtain the data we need.

4.2 System Simulation (to obtain the data)

Our system is composed of one dataset. A link to the CSV file can be found here.

4.3 Data set

Our data was previously collected using the Twitter API.

4.4 Data analysis

4.4.1 Expected results

Our expected results will be the topics most commonly discussed by Valencia citizens over time. The result will be shown as a collection of the top topics, as well as a graph to display their popularity over time.

4.4.2 Data analysis

Once our data is cleaned, we will use topic modeling in order to find the most common topics found in the dataset. We will use the study Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis[7] and a BERTopic tutorial by Martin Grootendorst [5] as references when conducting the analysis. Our results will allow the users to gain a better understanding of what issues are most important to Valencian citizens so that they can be more informed to take action and implement new policies.

BERTopic

There are many reasons for using the BERTopic algorithm in this project. BERTopic has been shown to outperform other traditional topic modeling algorithms, such as Latent Dirichlet Allocation (LDA), on several benchmark data sets. This is due to the use of the BERT model, which has been pre-trained on a large corpus of text and has shown exceptional performance in various natural language processing tasks. Furthermore, BERTopic allows for the customization of several parameters, such as the number of topics, the distance metric, and the clustering algorithm used. This flexibility makes it easier to tailor the topic modeling process to the specific needs of your research. Additionally, BERTopic is scalable and can handle large volumes of data, which is essential for our study when it comes to handling a big data set of tweets. These results will allow us to understand the most pressing topics that citizens are discussing [5].

5 Results

5.1 Data cleaning

To pre-process our data, we first removed duplicate entries and then used regex to remove URLs, extra whitespace, and non-alphanumeric characters. Additionally, we choose to filter out non-Spanish tweets; in order to do so, we used the "idioma" column of the original dataset as well as the Spacy Python library. We ended up with a CSV file containing 101,764 rows, meaning that we removed around 11,160 lines from the original dataset.

5.2 Topic Modeling

5.2.1 Extracting Topics

The Topic Modeling gave us the 10 most frequent topics from our data set. The -1 topic is referring to all the outliers meaning this one should be ignored. However, this is still one of the results and therefore will be shown in the table as well together with the other topics. The following ones are stated in the following table including the topic ID and the name of the topic, starting with the most frequent topic :

Topic	Name
-1	valencia hoy ms si
0	si voy ir valencia
1	si valenci quiero vida
2	partido jugadores equipo jugar
3	sanidad hospital pblica madrid
4	pp violadores deuda valncia
5	gracias enhorabuena abrazo muchas
6	quiera domingo barns nacho
7	tren metro bus transporte
8	amanecer noches denia buenas
9	publicar acaba foto espaa

5.2.2 Topic Word Scores

Note that each topic is made up of a collection of words and that the name of each topic is indicated by these words and not necessarily the name of the topic itself. As a result, the topic itself must be inferred by the keywords it consists of. Figure 2 displays the frequency of the words that make up each topic. For example, Topic 7, whose words include "tren," metro," "bus," "transporte," and "de," most certainly refers to public transportation in Valencia. On the other hand, it is clear that some topics such as Topics 0, 1, and 2, could be the result of commonly occurring words in the Spanish language and not an actual topic; this is an area in which our model could improve. Thus for each topic, we need to analyze the individual topic word scores in order to truly understand them. Ignoring topics that only contain common Spanish words, we can see that the following topics are of frequent discussion in our dataset:

- Valencia sporting events
- Public health and hospitals
- Violence in relation to the PP political party and the government of Valencia
- Expressing appreciation and thanks
- Public transportation
- Posting photos

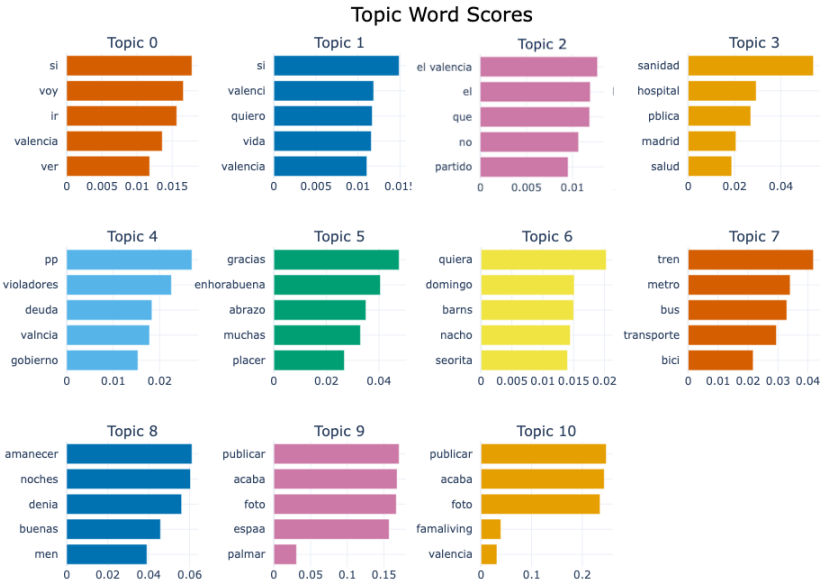

Fig. 2 Topic Word Scores

5.2.3 Topics over Time

Additionally, we analyzed the frequency of topics over time from October to December of 2022 in order to understand and visualize the most recent changes or fluctuations in discussed topics. This not only allows us to determine the most common topics discussed in the entire dataset but also those that have been most recently popular. The results can be shown in Figure 3, with a large spike in the number of tweets about topic 3 in the middle of November, and an increase in the number of tweets about Topic 4 in towards the end of November.

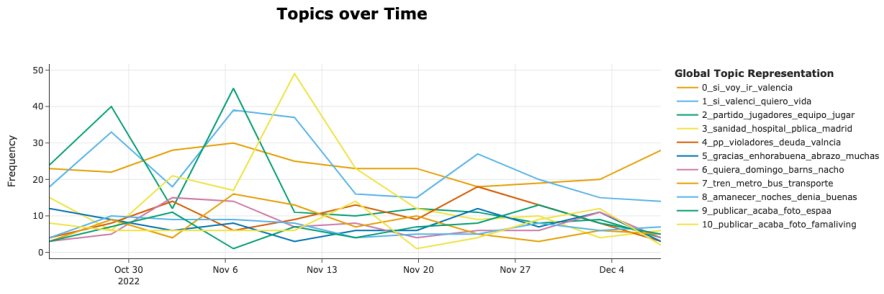

Fig. 3 Topics over Time

6 Discussion

6.1 Cleaned data set

We pre-processed our data set by removing the noisy parts to get a more clear result. Since the aim of this project is to understand what the biggest concerns are for the citizens in Valencia, we decided to proceed with the Spanish tweets and therefore removed the English ones from the data set. We believe that Tweets posted in the native language in Valencia were more likely to be posted by local citizens and not tourists, and would also best reflect their opinions. We also wanted to keep the dates for all of the Tweets in order to be able to show our results in a time-related graph. The URLs also added a lot of noise to our data since the words and numbers in the link wouldn't reflect or add any information to our topics; we would have seen the word "http" come up in many topics.

6.2 Topic modeling

Based on the results, topic modeling is clearly an effective way to understand and visualize what subjects are commonly discussed on Twitter. The results highlight how invested Valencia citizens are in their community; they not only care about public issues like health, transportation, and politics, but also about cultural events, expressing gratitude towards each other, and posting photos of their lives. Decision-makers can leverage this information to implement policies that address the public issues that are mentioned, as well as organize events or opportunities for citizens to express their appreciation of the community.

While our model did provide meaningful insight into the nature of citizens' discussions on Twitter, it could also have been improved to filter out more stopwords so that the results include fewer topics that just consist of common Spanish colloquial or filler words. This could be achieved by adding words like "si," "no," and "el" to the existing list of stopwords.

In the "Topics over Time" chart we could see a spike in Topic 4, which is about public health, coincides with the large strikes that occurred in November of that year in Madrid in defense of public health [8] explaining why "Madrid" is also in that topic's word collection. The majority of topics maintain a pretty steady frequency over this time period, indicating that Valencian interests or concerns do not fluctuate much from month to month apart from big current events.

7 Further Research

Further research could be done to implement sentiment analysis in order to see what topics might be the most pressing ones. By adding a sentiment analysis it is possible to determine what topics could be more negative and what that are more positive. We believe that the more negative ones could be more pressing and the biggest concerns of the citizens which also might be the ones decision-makers should prioritise.

Additionally, it would be interesting to see how topics develop over a longer time than what our results show now. The "Topics over time" chart only shows us the development of the topics over 3 months, but it could as well be interesting to see the development over 1 year or even longer. This information could be as important as showing the most recent topics. For instance for recurrent events such as Las Fallas in Valencia or other holidays in order to determine frequent concerns, and discover similarities and dissimilarities between the different years.

References

[1] Abalı, Gizem, Enis Karaarslan, Ali Hürriyetoğlu, and Feriştah Dalkılıç. "Detecting Citizen Problems and Their Locations Using Twitter Data." In 2018 6th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 30–33, 2018. https://doi.org/10.1109/SGCF.2018.8408936.

[2] Alotaibi, Shoayee, Rashid Mehmood, and Iyad Katib. "Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect." In 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), 330–35, 2019. https://doi.org/10.1109/FMEC.2019.8795331.

[3] Costa, Daniel G., Cristian Duran-Faundez, Daniel C. Andrade, João B. Rocha-Junior, and João Paulo Just Peixoto. "TwitterSensing: An Event-Based Approach for Wireless Sensor Networks Optimization Exploiting Social Media in Smart City Applications." Sensors 18, no. 4 (April 2018): 1080. https://doi.org/10.3390/s18041080.

[4] Egger, R. and Yu, J. (2022). "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts". Frontiers in Sociology, 7. https://doi.org/10.3389/fsoc.2022.886498.

[5] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". ArXiv:2203.05794, 2022. https://doi.org/10.48550/arXiv.2203.05794.

[6] Guo, Weisi, Neha Gupta, Ganna Pogrebna, and Stephen Jarvis. "Understanding Happiness in Cities Using Twitter: Jobs, Children, and Transport." In 2016 IEEE International Smart Cities Conference (ISC2), 1–7, 2016. https://doi.org/10.1109/ISC2.2016.7580790.

[7] Jang, Hyeju, Emily Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. "Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis." Journal of Medical Internet Research 23, no. 2 (February 10, 2021): e25431. https://doi.org/10.2196/25431.

[8] Li, Mengdi, Eugene Ch'ng, Alain Chong, and Simon See. "The New Eye of Smart City: Novel Citizen Sentiment Analysis in Twitter." In 2016 International Conference on Audio, Language and Image Processing (ICALIP), 557–62, 2016. https://doi.org/10.1109/ICALIP.2016.7846617.

[9] Ruhlandt, Robert Wilhelm Siegfried. "The Governance of Smart Cities: A Systematic Literature Review." Cities 81 (November 2018): 1–23. https://doi.org/10.1016/j.cities.2018.02.014.

[10] The Local Spain. "Thousands Rally in Defence of Madrid Public Healthcare," November 13, 2022. https://www.thelocal.es/20221113/roughly-200000-people-rally-in-defence-of-madrid-public-healthcare.

[11] "Twitter: Number of Users Worldwide 2024." Statista. Accessed May 5, 2023. https://www.statista.com/statistics/303681/twitter-users-worldwide.

[12] Zuo, Fan, Abdullah Kurkcu, Kaan Ozbay, and Jingqin Gao. "Crowdsourcing Incident Information for Emergency Response Using Open Data Sources in Smart Cities." Transportation Research Record 2672, no. 1 (December 1, 2018): 198–208. https://doi.org/10.1177/0361198118798736.