Homelessness in California

Nicola Bee[†], Farshad Jafarpour [‡], Lorenzo Saccaro[⋆], Menglu Tao[◇]

***Abstract*—In recent years there has been a surge in the number of homeless people in the United States causing a social crisis in many large cities. This report explores the issue of homelessness in California through the lens of network science and natural language processing. By building word and hashtag networks starting from a collection of tweets and analyzing their properties such as degree distribution, robustness, and the presence of communities, we gather insights into this complex phenomenon. Moreover, by performing topic modeling with BERTopic and some simple sentiment analysis we are able to determine which are the characterizing themes connected to the homeless crisis, such as the lack of affordable housing, drug abuse, and mental health issues, and we explore what are the public opinions and feelings.**

I. Introduction

Homelessness has become an increasingly prevalent issue in many urban areas of the United States, and California is no exception. The state has experienced a surge in homelessness in recent years, with an estimated 161,000 people experiencing homelessness on any given night, accounting for nearly half of the total homeless population in the country [1]. The problem is particularly acute in cities such as Los Angeles, San Francisco, and San Diego, where high housing costs and a shortage of affordable housing have contributed to the crisis. The causes of homelessness are complex and multifaceted, and include factors such as poverty, unemployment, mental illness, addiction, and domestic violence [2]. In California, rising housing costs and stagnant wages have made it difficult for low-income individuals and families to find affordable housing, leading to increased rates of eviction and homelessness [3]. Efforts to address homelessness in California have included the provision of emergency shelter, the construction of affordable housing, and the implementation of supportive housing programs that offer a range of services, including counseling, job training, and healthcare, to help individuals and families overcome the challenges that contributed to their homelessness [4].

By adopting network science tools on a collection of tweets we aim to gather insights on the complex problem of homelessness in California based on what the public's discussion is about. We aim to answer some research questions such as: what are the causes of this phenomenon? What is being done to solve the crisis? What is the public perception and what are people's opinions on this subject?

Department of Physics and Astronomy "Galileo Galilei", University of Padova
[†]nicola.bee.1@studenti.unipd.it
[‡]farshad.jafarpour@studenti.unipd.it
[⋆]lorenzo.saccaro@studenti.unipd.it
Department of Information Engineering, University of Padova
[◇]menglu.tao@studenti.unipd.it

As a sanity check and to compare our findings we perform the same analysis for New York State with a major focus on New York City. This decision seems natural since New York is the most populous city in the United States. Moreover, this comparison arises also from the data collected on California.

This work is structured as follows: in Section II we describe the data gathering and preprocessing phases, in Section III we dive into the analysis of the major networks metrics, the community detection, network robustness and more. In Section IV we perform topic modeling to extract relevant themes and in Section V a simple sentiment analysis is conducted on the gathered tweets to investigate people's overall feelings or attitudes towards the homeless crisis.

II. Data gathering and preprocessing

For our data collection in this project, we utilize the Twitter API to gather tweets over a period of about 30 days, starting Jan 5, 2023. The Twitter API [5] provides a powerful tool to easily filter and collect a large amount of data within our specified timeframe and search criteria.

A first attempt is made using only selected hashtags with the help of Google Trends [6], but the number of obtained tweets is insufficient to perform any meaningful analysis. Since there are no trending hashtags related to our research topic, we opt to create a query that combines keywords such as 'homeless', 'homelessness', 'unhoused', etc... with location markers. Therefore, our search terms include 'Califonia' and cities such as San Francisco, Los Angeles, Oakland, and San Jose, as well as their abbreviations such as CA, SF, and LA. The same approach is followed to retrieve tweets about New York State and City. The two datasets contain ≈ 90000 and ≈ 27000 tweets respectively.

The data preprocessing phase consists of several steps to prepare the collected data for analysis. Firstly, we remove all emojis from the text to ensure that only text-based data is retained. We also filter the tweets to only include those in English, as our analysis focuses on this language. Duplicates are also dropped in this phase. In order to clean the text and extract words we adopt the Natural Language Toolkit for Python (`nltk` [7]). In particular, after removing punctuation, HTTP links, mentions, and stop words, tweets are tokenized and lemmatized.

Extracted words and, separately, hashtags are then connected if they appear in the same tweet: if the pair appears multiple times the corresponding edge weight is increased accordingly. From each dataset, we obtain a word and a hashtag network with weighted edges that are saved as an edgelist file for later analysis.

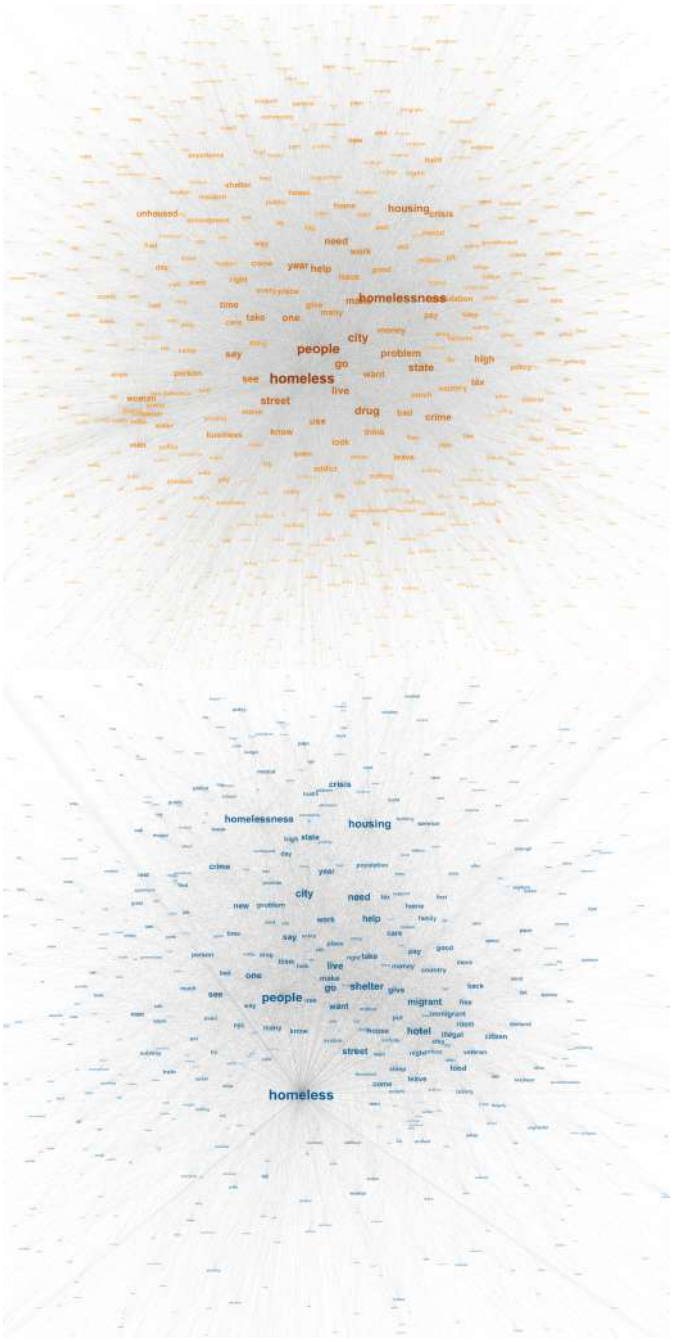


Fig. 1: Words network visualization. California is on top, and New York is on the bottom.

III. Network analysis

To visualize the networks we adopt Gephi [8], an open-source network analysis and visualization software package. In particular, we use the ForceAtlas2 [9] layout algorithm which is based on an attraction-repulsion force between nodes. Moreover, we highlight nodes with increasing color intensity and text size based on their degree, while edges are more marked depending on their weight. The plots for the word and hashtag networks are shown in Fig. 1 and in Fig. 2.

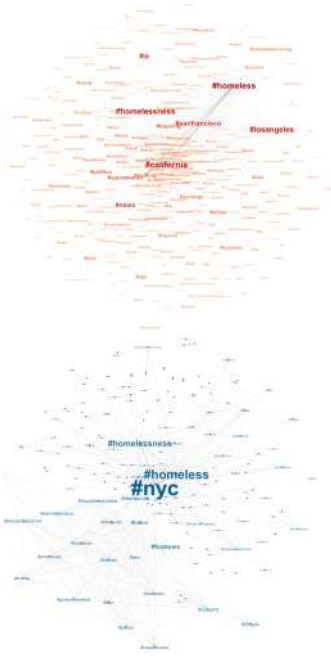


Fig. 2: Hashtag network visualization. California is on top, and New York is on the bottom.

	Words		Hashtag	
	CA	NY	CA	NY
Nodes	2330	1299	463	175
Edges	42377	12221	2433	645
Avg. degree	36.38	18.82	10.51	7.37
Avg. squared degree	11388	3083	508	179
Avg. weighted degree	543.39	175.17	25.05	13.38
Avg. weighted squared degree	9954517	825898	9825	934
Density	0.0156	0.0145	0.0227	0.0424
Diameter	4	5	∞ (6)	∞ (5)
Avg. path length	2.03	2.13	∞ (2.49)	∞ (2.33)
Weighted diameter	56	43	∞ (9)	∞ (9)
Weighted avg. path length	16.21	13.19	∞ (3.30)	∞ (3.44)

TABLE 1: Major network metrics

A. Network Metrics Analysis

By analyzing different metrics, we can get insights into the structure, behavior, and importance of the nodes and edges within the network. The following quantities are investigated:

- **Network structure**: In Tab. 1 we report metrics that describe some of the basic aspects of a network such as the number of nodes and edges, the average degree, diameter, and average path length. The lower number of tweets in the New York dataset translates into smaller word and hashtag networks. It is interesting to observe that both hashtag networks have disconnected components: for this reason, we report in parentheses the value of the diameter and the average path length for the giant component.
- **Authorities and Hubs (HITS)**: HITS (Hyperlink-Induced Topic Search) is a link analysis algorithm used to identify relevant web pages for a given query. It is based on the assumption that a good page is one that is both authoritative and hub-like, meaning it has many links to other relevant pages and is linked to by many

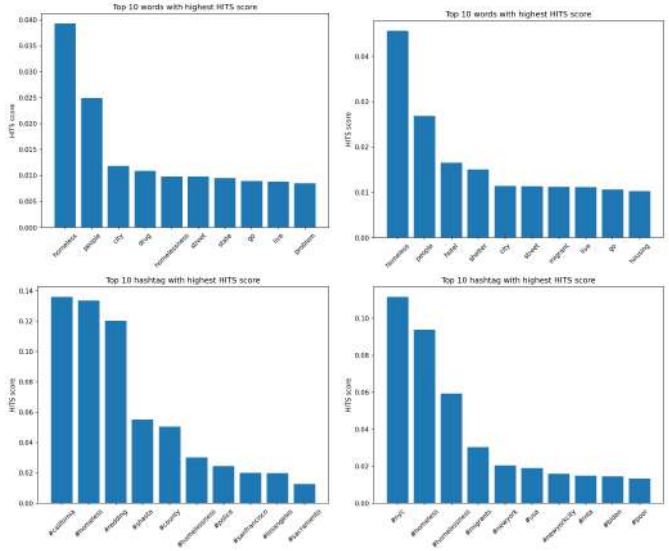


Fig. 3: Top 10 words and hashtags with highest HITS scores for California (left) and New York (right) networks

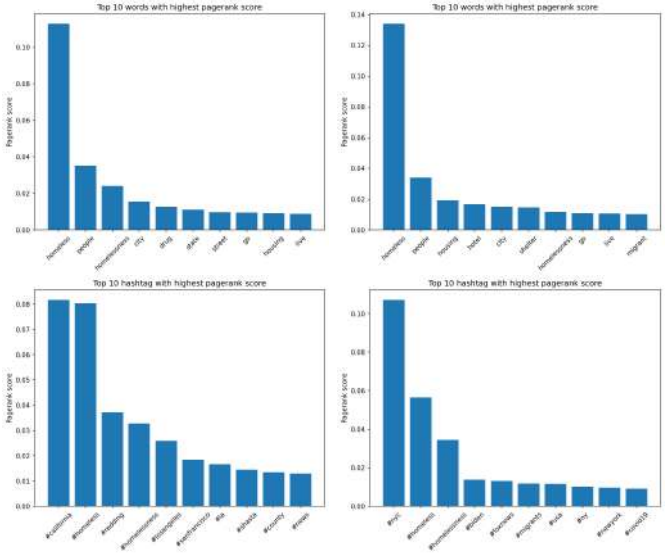


Fig. 4: Top 10 words and hashtags with highest PageRank scores for California (left) and New York (right) networks

other relevant pages. The algorithm works by iteratively calculating two scores for each page: an authority score and a hub score. Pages with high authority scores are seen as more relevant to the query, while pages with high hub scores are seen as good sources of information. HITS is useful for ranking web pages and identifying relevant content.

Fig. 3 shows the top 10 words and hashtags with the highest HITS score(since we are dealing with undirected networks, the hub and authority scores are the same). Among the highest 10 words, the results are quite similar, words like homeless, people, street, live, and city appeared in both networks, suggesting that these are important and pervasive issues related to homelessness that are being discussed in both regions, while hotel, migrant, and shelter are specific to New York and drug to California As for hashtags, the interesting thing is the first hashtag both are names of locations. In the California hashtag network, it's #redding, which is the name of a city in Northern California. This suggests that discussions of homelessness in California are not just focused on the large metropolitan areas like Los Angeles and San Francisco, but also on smaller cities and towns. On the other hand, the top hashtag in the New York network is #nyc, which is not surprising given that New York City is one of the largest and most densely populated cities in the world. The presence of hashtags related to public transportation #mta and political figures #biden suggest that discussions of homelessness in NYC may also touch upon issues related to government policies and social welfare programs.

- **Page rank**: a link analysis algorithm used by search engines to rank websites in their search engine results. It works by assigning a score to each webpage based on the quantity and quality of other webpages linking to it. The more incoming links a webpage has from other high-quality webpages, the higher its PageRank score. PageRank helps to identify the most important and relevant webpages for a given search query. Its advantage lies in its ability to provide a more accurate and unbiased ranking of webpages, as it is not influenced by paid advertising or other manipulative tactics.

  Fig. 4 shows the top 10 words/hashtags with the highest page rank score. For words, the result is similar to the HITS one, the two networks share quite a lot of common words. For hashtags, the result is slightly different from what we get from the HITS score. In the California hashtag network, 4 out of 10 hashtags are location: #califonia, #redding, #losangeles, #sanfrancisco. This indicates those locations are involved with serious homeless situations. And in the New York hashtag network, hashtags become more political, involving #ericadams, who is the mayor of new york city, #biden, the current president of the USA. This is the reflection of local citizens relating the homelessness situation with politicians.
- **Closeness centrality, harmonic centrality and betweenness centrality**: Closeness centrality measures how easily a node in a network can reach all other nodes, taking into account the length of the shortest paths between them. Nodes with high closeness centrality are well connected and can quickly disseminate information throughout the network.

  Harmonic centrality is a variation of closeness centrality that considers the sum of the reciprocal distances between a node and all other nodes. It emphasizes the importance of nodes that are close to many other nodes and penalizes nodes that are far away from the rest of the network.

  Betweenness centrality quantifies the importance of a node in the flow of information or resources between other nodes. It measures how often a node lies on the

Closeness	Betweenness	Harmonic
homeless	homeless	homeless
people	people	people
homelessness	homelessness	homelessness
city	city	city
drug	state	drug
state	housing	state
go	drug	go
street	blah	street
housing	live	housing
live	go	live

TABLE 2: Top 10 words by closeness centrality, betweenness centrality and harmonic centrality of the California network

Closeness	Betweenness	Harmonic
homeless	homeless	homeless
people	people	people
housing	housing	housing
city	city	city
shelter	homelessness	shelter
hotel	shelter	hotel
homelessness	hotel	homelessness
go	live	go
need	go	need
live	crisis	live

TABLE 3: Top 10 words by closeness centrality, betweenness centrality and harmonic centrality of the New York network

shortest paths between all other pairs of nodes in the network. Nodes with high betweenness centrality act as bridges or bottlenecks in the network and are essential for efficient communication and resource allocation. Tab. 2 and Tab. 3 display the 10 words with the highest scores of closeness, harmonica, and betweenness centrality in the California and New York word network respectively. In both networks, the top three words by all three centrality measures are **homeless**, **people**, and **city**, indicating a focus on the issue of homelessness in urban areas and the impact of homelessness on individuals and communities. Additionally, words like **housing** and **live** appear in both networks, suggesting a focus on the availability and quality of housing, and the lived experiences of homeless individuals.

However, there are also some differences between the two networks. In the California network, words like "drug", "state" and "go" appear among the top 10, suggesting a focus on the role of the state government in addressing homelessness and the need for action and solutions, while underlying the drug abuse by homeless people. In contrast, in the New York network, words like **shelter**, **hotel**, and **need** appear in the top 10, suggesting a focus on the availability of temporary housing options and the basic needs of homeless individuals.

Tab. 4 reports the 10 hashtags with the highest scores of closeness, harmonica, and betweenness centrality in the California hashtag network. The hashtags with the highest scores in closeness centrality and harmonic centrality are similar, but hashtags with high betweenness scores are different from them. The same situation appears in Tab. 5 which shows the 10 hashtags with the highest scores in the New York network.

B. Degree distribution and gamma estimation

The degree distribution is an important property of networks that provides insight into the connectivity and structure of a network. In particular, it is interesting to determine if a network is scale-free, i.e. its degree distribution follows a power law $p_k = Ck^{-\gamma}$, and estimate the degree exponent γ. This is not a trivial task and some key steps need to be followed with care. More in detail, we follow the suggestions

Closeness	Betweenness	Harmonic
#california	#homelessness	#california
#homeless	#homeless	#homeless
#homelessness	#california	#homelessness
#losangeles	#losangeles	#losangeles
#sanfrancisco	#la	#la
#la	#sanfrancisco	#sanfrancisco
#news	#news	#news
#sacramento	#twitterblue	#sacramento
#crime	#sandiego	#housing
#housing	#nyc	#crime

TABLE 4: Top 10 hashtags by closeness centrality, betweenness centrality and harmonic centrality of the California network

of [10] and we adopt the `powerlaw` package [11], which is suited for the analysis of heavy-tailed distributions.

Firstly, the frequency of the degree is plotted using a log-log scale and the empirical probability mass function is obtained using a log binning. Before fitting the distribution, it is important to determine k_{min}, which is the small degree cutoff that separates the saturation region from the interval of degrees where the power law is a valid approximation. The `powerlaw` package automatically tests multiple values of k_{min} and returns the one that results in the lowest Kolmogorov-Smirnov distance (D) between the fitted distribution and the real one. It could happen that multiple k_{min} have close corresponding D values: in that case, we opt for the

Closeness	Betweenness	Harmonic
#nyc	#nyc	#nyc
#homeless	#homelessness	#homeless
#homelessness	#homeless	#homelessness
#foxnews	#ny	#foxnews
#democrats	#foxnews	#democrats
#illegalimmigrants	#democrats	#biden
#biden	#illegals	#democrat
#hope	#democrat	#blacklivesmatter
#migrants	#affordablehousing	#covid19
#bidenbordercrisis	#ericadams	#illegals

TABLE 5: Top 10 hashtags by closeness centrality, betweenness centrality and harmonic centrality of the New York network

	Words		Hashtag	
	CA	NY	CA	NY
γ	2.09 $\pm$ 0.08	2.51 $\pm$ 0.21	2.27 $\pm$ 0.12	2.12 $\pm$ 0.16
D	0.21	0.35	0.16	0.23
p-value	0.808	0.343	0.972	0.898
exponential	+	+	++	+
stretched exponential	+	+	+	+
truncated power law	-	+	+	-
lognormal positive	-	+	+	+

TABLE 6: Fit estimates of gamma exponent. The Kolmogorov-Smirnov statistic (D) and the corresponding p-value are also reported. The sign of the Loglikelihood ratio between the power law and other crossover distributions is repeated if the result is statistically significant (based on the corresponding p-value of the Loglikelihood ratio).

lowest one, which allows for a larger fit region. Once k_min is fixed, the power law fit returns the γ value. The results are shown in Tab. 6 and the corresponding plots are in Fig. 5.

All four networks seem to display an ultra-small world behavior since $2 < \gamma < 3$, however, the uncertainties on the estimated γ values are pretty high. Moreover, specifically for the New York word network, we observe a very high k_{min} and a p-value that, while being above the 0.05 standard threshold, is not very reassuring. However, there are no statistical indications that the scale-free nature of these networks should be rejected.

Finally, one should consider other heavy-tailed distributions (also called crossover) that may be a more suitable explanation for the measured pmf, such as the exponential, stretched exponential, truncated power law, and the lognormal positive. The `powerlaw` package computes the Loglikelihood ratio between the fitted power law and the aforementioned distributions. A positive value suggests that the power law is a more accurate representation of the real pmf, while a negative value favors the alternative hypothesis. The ratio is also paired with a corresponding p-value that can be used to determine the statistical significance of the obtained result. In Tab. 6 we report the sign of the ratios and we repeat it if its value is statistically significant. This only happens for the exponential distribution in the California hashtag network, which can therefore be excluded. In general, the power law is the better explanation, the only exception are the truncated power law for the California word and New York hashtag networks and the lognormal positive for the California word network. In all cases, however, there is no statistical evidence to reject the power law fit.

C. Assortativity

Assortativity refers to the tendency of nodes in a network to be connected to other nodes that have similar characteristics or properties. In other words, it describes the degree to which nodes with similar attributes tend to be connected to each other within a network. In this work, we focus on degree assortativity.

There are two quantities that can be used to measure it: the assortativity coefficient μ and Pearson's correlation coefficient r. A positive value indicates that nodes tend to be connected to

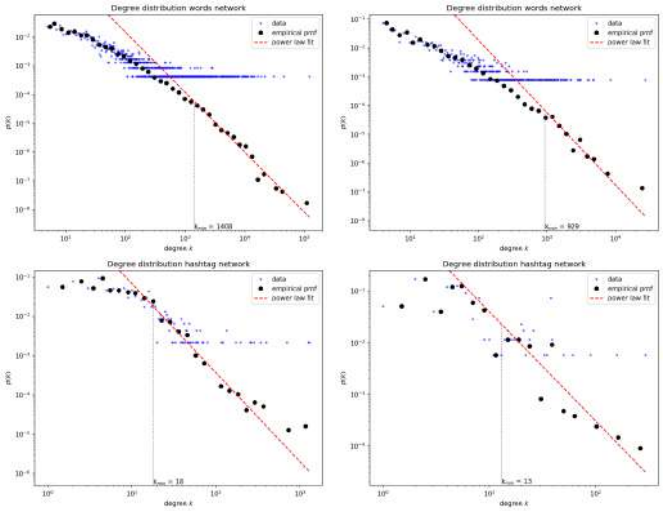


Fig. 5: Degree distribution in a log-log plot. The black dots represent the empirical probability mass function (pmf) obtained with a log binning. In red we show the power law fit: $p_k = C\, k^{-\gamma}$. California results are on the left, and New York on the right.

	Words		Hashtag	
	CA	NY	CA	NY
μ	-0.212 $\pm$ 0.004	-0.239 $\pm$ 0.007	0.05 $\pm$ 0.02	-0.01 $\pm$ 0.04
r	-0.1237	-0.1430	-0.1641	-0.2006
R-S μ	-0.068 $\pm$ 0.005	-0.03 $\pm$ 0.02	-0.25 $\pm$ 0.04	-0.19 $\pm$ 0.06
R-S r	-0.0507	-0.0788	-0.1728	-0.2223

TABLE 7: Assortativity coefficient μ extracted from the fit and Pearson's correlation coefficient r for the real network and the degree preserving randomization with simple links (R-S)

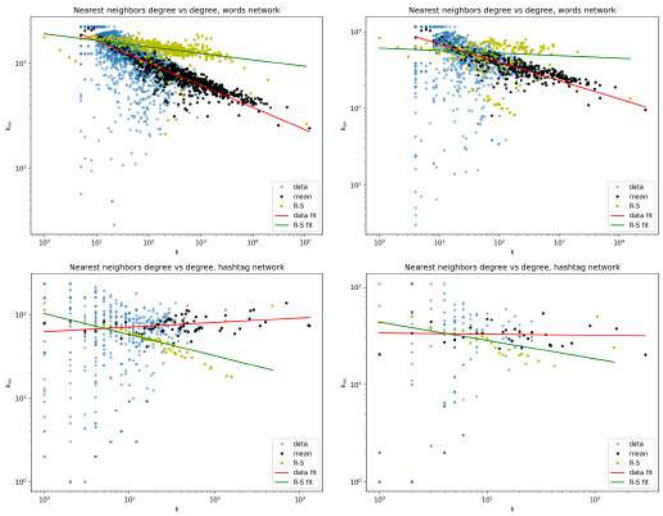


Fig. 6: Average neighbor degree k_{nn} vs degree k. The average value for each k is shown in black. The linear fit in the log-log plot is in red, while the data and the corresponding fit for the degree preserving randomization with simple links (R-S) are in green. California results are on the left and New York on the right.

	Words		Hashtag	
	CA	NY	CA	NY
$\boldsymbol{\kappa}$	18319	4714.84	392.21	69.82
f_c	0.9999	0.9998	0.9974	0.9855

TABLE 8: Inhomogeneity ratio ($\boldsymbol{\kappa}$) and breaking point (f_c)

others with similar degrees, while a negative value indicates the opposite. Pearson's correlation coefficient r, which lies in the [-1,1] interval, is computed directly by applying an algorithm from the NetworkX package [12]. To compute the assortativity coefficient μ, we plot the average neighbor degree k_{nn} for each node vs its degree, then take the average value for each degree k. We extract μ as the slope of the linear fit in the log-log scale. The obtained values are reported in Tab. 7 while the corresponding plots are in Fig. 6.

California and New York word networks show a clear presence of disassortative behavior by both metrics. For the hashtag ones, the μ measure suggests neutrality while the r significantly differs from zero. However, the value of μ is to be trusted more since there is only a correlation with the sign of r and the scale-free nature of the hashtag network has not been excluded as described before.

To determine if the disassortativity present in the word networks is structural ($\gamma < 3$), i.e. it depends on the degree distribution of the network, we perform a degree preserving randomization with simple links (R-S), leveraging the NetworkX function `double_edge_swap` [13] and setting the `nswap` parameter to 10 times the number of edges. The corresponding μ and r values are extracted following the same procedure of the real networks and the results are in Tab. 7 and in Fig. 6.

Both the assortativity coefficient and Pearson's correlation coefficient suggest that the disassortativity present in the word networks can only partially be explained by their degree distribution. There are deeper reasons that go beyond the structure of the network. For the hashtag ones the randomization seems to suggest that some structural disassortativity should be present due to the scale-free nature of the network, therefore there could be other, non-random, explanations for the observed neutrality.

D. Network Robustness

Network robustness refers to the ability of a network to maintain its functionality and performance in the face of different types of disturbances, failures, or attacks. In this work, we perform an analysis of both the bipartite network and its projection resilience. In particular, we focus on two main types of nodes failure:

- **Random failures**: a fraction of the nodes of the network is randomly selected and removed.
- **Attacks**: nodes with the highest degree, i.e. hubs, are targeted first.

In Tab. 8 we compute the inhomogeneity ratio, $\boldsymbol{\kappa} = \langle k^2 \rangle / \langle k \rangle$ and the breaking point $f_c = 1 - 1/(\boldsymbol{\kappa} - 1)$.

In Fig. 7 we plot the giant component (GC) fraction, which is defined as the fraction of nodes in a network that belongs

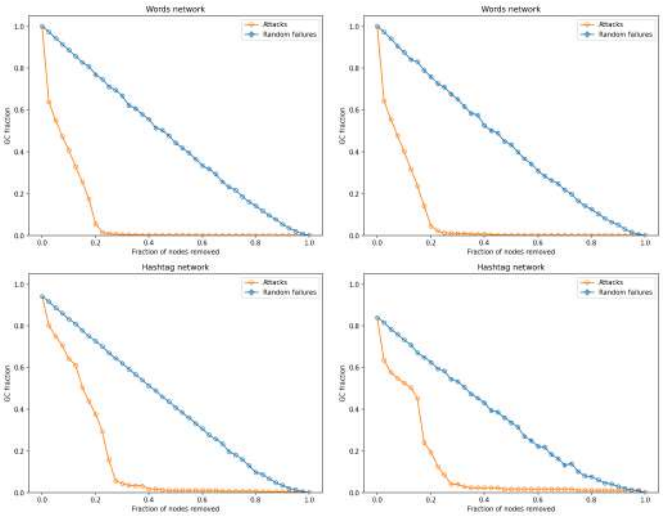


Fig. 7: Gian component (GC) fraction as a function of the percentage of removed nodes. For each network (California - left, New York - right) random failure results are shown in blue while attacks are in orange.

to its largest connected component, as a function of the percentage of removed nodes. For the random failure, given its stochastic nature, the test is repeated 100 times and the average result with its corresponding error bar is reported.

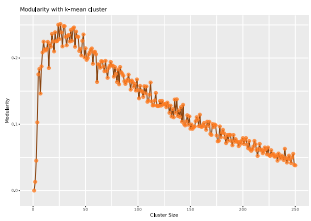
As expected, due to their free-scale nature, all four networks have a very high breaking point with a linear decrease of the GC fraction when dealing with random failures, requiring the removal of almost all nodes to break down the network. Conversely, they are particularly vulnerable to attacks: with only slightly above 20% of the highest degree nodes removed, the word networks collapse while the hashtag ones barely survive up to 60-70%. In both cases, the decrease in the GC fraction is extremely fast.

E. Community Detection

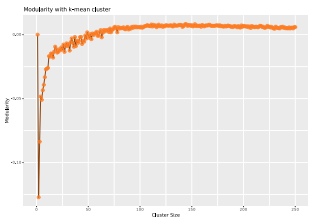
Community detection aims to identify densely connected groups or clusters of nodes in a network. These groups should help understand the structure of the network and provide valuable insights into its underlying properties and patterns. We first consider some measures for cluster quality, namely modularity, and conductance.

Modularity: Modularity is a widespread measure of the quality of a cluster. It represents the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. The value for unweighted and undirected graphs lies in the range [-½, 1]. As stated before it is also exploited by two of the previous algorithm (*Fast greedy* and *Louvian*). We can study the modularity with respect to the number of clusters by performing iteratively the k-means clustering on the network changing the number of centers, as shown in Fig. 8.

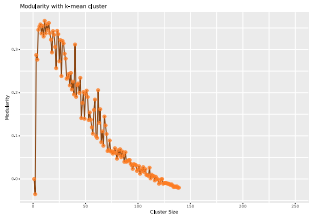
It can be seen that hashtags and words have two different behavior:



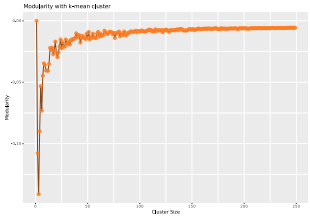
(a) California - hashtags



(b) California - words



(c) New York - hashtags



(d) New York - words

Fig. 8: Modularity scores of the networks

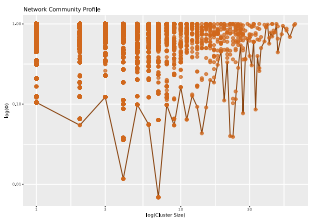
- The *hashtags* networks show a positive spike around the region from 1 to 25 clusters. This is a standard trend in networks and the maximum of the modularity provides an estimate of which the optimal number of clusters is. We can notice that the NY network reaches a value higher than 0.35, which suggests a significant community structure, whereas the CA network reaches a lower value, around 0.25
- The *words* networks instead have a negative spike and subsequently slowly settle down to a small negative value. This latter trend suggests that the graphs are almost fully connected or present strongly overlapping communities, and breaking a graph into smaller clusters would break a lot of connections.

Conductance: Conductance is another measure of the quality of a cluster. Specifically, it measures the ratio of the number of edges that cross the boundary between two sets of nodes, to the total number of edges that are incident to those two sets. A network with high conductance implies that there is high interconnectivity and communication between different groups of nodes in the network, whereas low conductance represent an isolated cluster with few outward connections. By running the k-means algorithm with a different value of k, for each community found we can insert a point in a graph showing the conductance value versus the number of nodes in the cluster. The minimum of this graph is called *network community profile* and is shown in Fig. 9 for the hashtags networks.

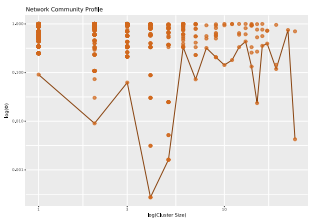
Wherever a local minimum is found this indicates that that cluster has good performance. Note that, unlike modularity, we are evaluating here how good individual clusters are and not the optimal number of clusters to use.

Clustering analysis: Different cluster algorithms have been tried, in particular:

- **Fast greedy modularity optimization**: a widely used algorithm for detecting communities in complex networks. Its main advantage lies in its efficiency, as it can quickly identify large communities in networks with millions of nodes. The algorithm works by greedily optimizing the modularity score at each step by merging the pair of communities that leads to the largest increase in modularity. This process continues until no further increase in modularity is possible.
- **Spectral clustering**: another popular algorithm for community detection that leverages the spectral properties of the network's adjacency matrix. It works by first computing the eigenvalues and eigenvectors of the matrix and then using them to perform clustering. The algorithm maps the nodes of the network onto a lower-dimensional space defined by the eigenvectors, where nodes that are likely to belong to the same community are close together. In this way, it can handle networks with complex community structures, and it is more stable and less expensive than recursive bisection methods.
- **Betweenness clustering**: a community detection algorithm that is based on the betweenness centrality measure. It works by first computing the betweenness of all edges in the network and then iteratively removing the edges with the highest score until the network breaks into disconnected components that correspond to the communities in the network. This algorithm has several advantages, including its simplicity and its ability to detect overlapping communities. However, its scalability to very large networks is limited, as computing betweenness centrality for all edges can be computationally expensive.
- **Louvain clustering:** another modularity-based algorithm that works by optimizing the modularity score of the network. The algorithm starts by assigning each node to a separate community, and then iteratively optimizes the modularity score by moving nodes between communities. At each iteration, nodes are moved to the community which results in the highest increase in modularity, and the process continues until no further improvement in modularity is possible. The Louvain algorithm is highly scalable, can handle very large networks, and is also capable of detecting overlapping communities.



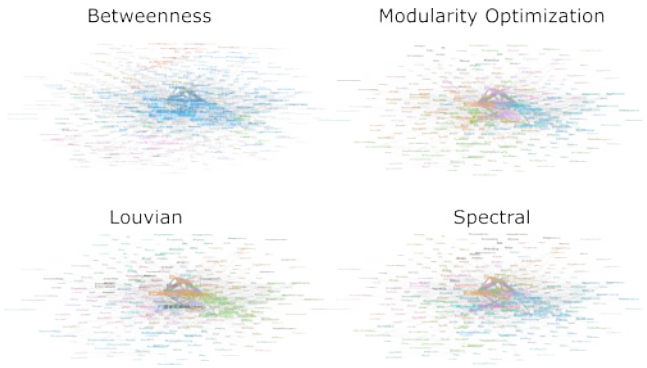
(a) California



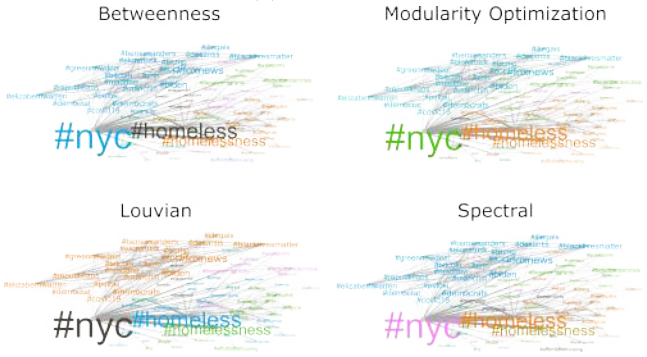
(b) New York

Fig. 9: Conductance and network community profile of hashtags networks

These algorithms are performed on both California and New York datasets, using hashtags and word networks for a total of four different graphs. As we have seen the word networks have poorer performances in modularity and this results in less useful cluster information. For this reason, the related



(a) California



(b) New York

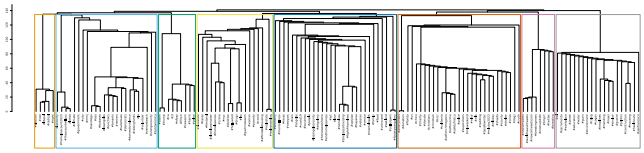
Fig. 10: Different community detection plots for hashtags network

plots are not presented. Regarding the hashtag networks we can see in Fig. 11 that although the overall structure is similar, different algorithms display different behaviours. In particular we can observe that betweenness tends to produce many more clusters than the other algorithms: in California's network, for example, betweenness produces 138 clusters, mainly of isolated nodes, while the others all ends with 11 or 12 clusters. Futhermore we can observe that, in New York graphs, communities tend to stay similar as the algorithm changes. This reflect the higher modularity of the network respect to California's one and suggest that its communities are more compact and less overlapping. In fact in New York network, looking for example at the Louvian's communities we can identify some topics in the clusters, for example one about foregin countries (with hashtags like *#gaza, #nigeria, #iran*), one more focused on immigration (*#illegalimmigration, #bidenbordercrisis, #american*), and one about politics (*#trump, #biden, #foxnews, #blackslivematter*).

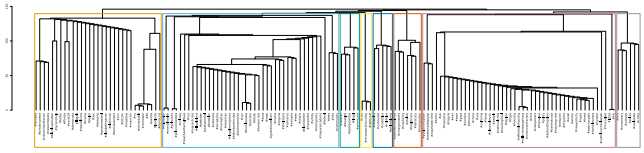
For the Fast greedy modularity optimization, we can also represent in Fig. 11 the dendrogram to show in which order different hashtags have been divided into communities.

IV. Topic Modeling

In order to extract topics from our tweets datasets, we adopted BERTopic [14] a topic modeling technique that leverages the combination of a deep neural network to create



(a) California



(b) New York

Fig. 11: Dendrograms of hashtags network

embeddings for text data and clustering algorithms to group similar documents[1] into topics.

More in detail, the overall pipeline consists of 5 major steps:

- **Preprocessing**: before the creation of the embeddings the text needs to be cleaned: in general, not much preprocessing is required. In our case links, mentions, and hashtags are removed from the tweets while duplicates and non-English ones are dropped.
- **Document emdeddings**: a pre-trained Sentence Transformer model is used to generate the document embeddings. Since the size of the datasets is limited and we have access to a GPU to speed computations, we opt to use the largest (but slowest) model available for the English language which provides the highest quality outputs: `all-mpnet-base-v2` [15].
- **Dimensionality reduction**: the dimensionality of the document embeddings can be quite high, which can make clustering computationally expensive and slow. To overcome this, dimensionality reduction techniques are applied to reduce the dimensionality of the embeddings while preserving their semantic content. BERTopic uses UMAP [16] (Uniform Manifold Approximation and Projection), a non-linear dimensionality reduction technique that is well-suited for high-dimensional data. UMAP is used to reduce the dimensionality of the embeddings to a lower dimensional space while preserving their pairwise distances. The algorithm has two main parameters that require tuning. The first is `n_neighbors`, which is the number of neighboring sample points used when making the manifold approximation and balances between the local and global view of the embedding structure. The second is `n_components`, which refers to the dimensionality of the embeddings after reducing them. The idea is to reduce dimensionality as much as possible to ease the job of the clustering algorithm whilst trying to maximize the information kept in the resulting embeddings.

[1] tweets in our case

- **Clustering**: similar documents are grouped into topics. BERTopic uses hierarchical clustering with the HDBSCAN [17] algorithm, which is a density-based clustering algorithm that can automatically determine the number of clusters and is robust to noise and outliers. HDBSCAN creates a hierarchy of clusters, with each cluster being a subset of its parent cluster. The clusters are created by grouping together embeddings that are close to each other in the reduced-dimensional space. The HBSCAN algorithm has two main parameters that require tuning. The first is `min_cluster_size`, which controls the minimum size of a cluster and thereby the number of clusters that will be generated. Larger values will result in fewer clusters of larger sizes. The second is `min_samples`, which is related to the number of outliers that are generated. Setting it to values lower than the one of `min_cluster_size` will reduce the amount of noise.
- **Topic selection and representation**: BERTopic selects the most representative documents from each topic to help users understand the content of each topic and identify the most relevant documents within each topic. BERTopic uses a combination of techniques, such as c-TF-IDF [18] (Class-based Term Frequency-Inverse Document Frequency) and sentence similarity, to select the most representative documents. Moreover to reduce the redundancy and improve the diversity of keywords an algorithm called Maximal Marginal Relevance (MMR [19]) is implemented. MMR considers the similarity of keywords/keyphrases with the document, along with the similarity of already selected keywords and keyphrases. This results in a selection of keywords that maximize their diversity with respect to the document. Setting the diversity coefficient to 0.5 produces better results. Finally, stop words are removed from the topic representation, and using an `n_gram_range` of 1-2 allows having, for example, "New York" as a keyphrase instead of "New" and "York" as two separate keywords.

A schema of the processing pipeline is shown in Fig. 12

After training the model on our datasets, the number of generated topics is pretty high, and most are redundant. The BERTopic model implements a function, `reduce_topics`, that can automatically reduce the number of topics using HDBSCAN, or alternatively, one can set the desired target number of topics. Using the first approach we reduce the number of topics to 38 and 40 for the California and New York datasets respectively.

We also notice that a pretty large part of the tweets is classified as outliers: BERTopic allows reducing them using the `reduce_outlier` function. Multiple strategies are available:

- **Probabilities**: it uses the soft-clustering as performed by HDBSCAN to find the best matching topic for each outlier document.
- **Topic distributions**: this strategy uses the topic distribution to find the most frequent topic in each outlier document.
- **c-TF-IDF**: it calculates the c-TF-IDF representation for each outlier document and finds the best matching c-TF-IDF topic representation using cosine similarity
- **Embeddings**: using the embeddings of each outlier documents, it finds the best matching topic embedding using cosine similarity.

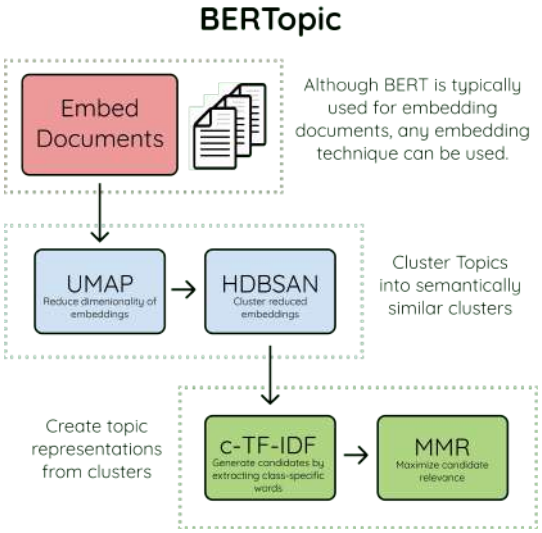


Fig. 12: Schema of BERTopic processing pipeline

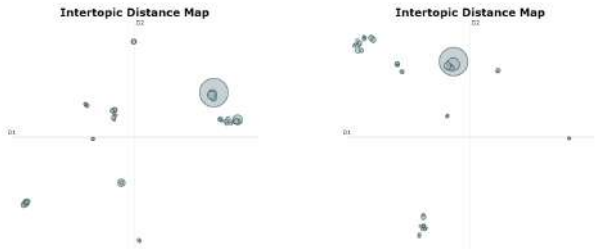


Fig. 13: 2D intertopic distance map: California's topics are on the left, New York's ones on the right

After some experimentation, the c-TF-IDF strategy is the one that leads to an equilibrated assignment of most of the outliers.

Before analyzing the extracted topics, we showcase some useful visualizations that help grasp the overall structure and geometry of our datasets. In Fig. 13 topics are positioned in a 2D plane according to their c-TF-IDF representation: closer topics have similar representation, while the size of the circle is proportional to the number of tweets assigned to that topic. For both California and New York datasets, topics tend to form clusters: it is easy to spot 9 and 7 of them respectively.

The similarity matrix in Fig. 14 is computed by applying cosine similarity to topic embeddings. Setting the `n_cluster` parameter, topics are ordered by their similarity. This results in blocks forming in the heatmap indicating which clusters of topics are similar. This block-like structure is visible in both

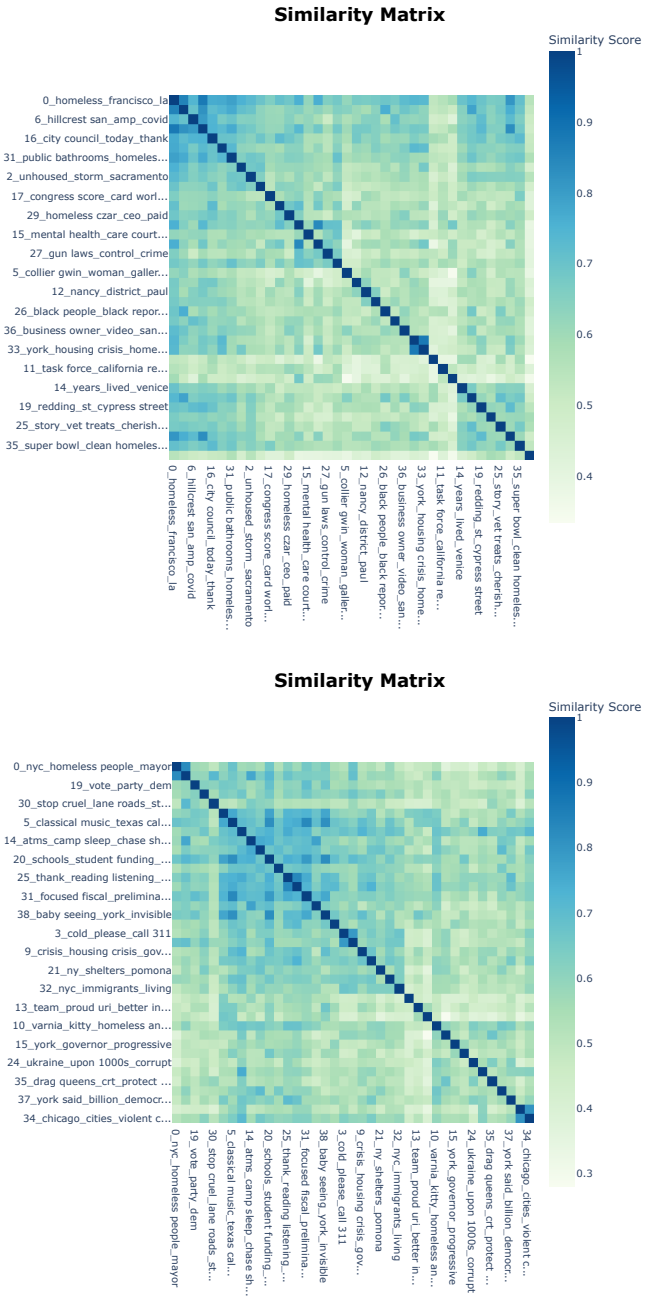


Fig. 14: Topics similarity matrix: California's is on top, New York's on bottom

matrices.

To further understand the structure of the extracted topics: in Fig. 15 we plot the hierarchical clustering to visualize how groups of topics are related to one another. This visualization can be leveraged to hierarchically merge topics if one wants to furthermore reduce their number. California's plot shows a finer structure where ramifications happen at a higher level wrt New York's. Moreover, it is simpler to understand why topics are grouped together in Califonia's hierarchy while in New York's there's a large cluster (in red) that is hard to make sense

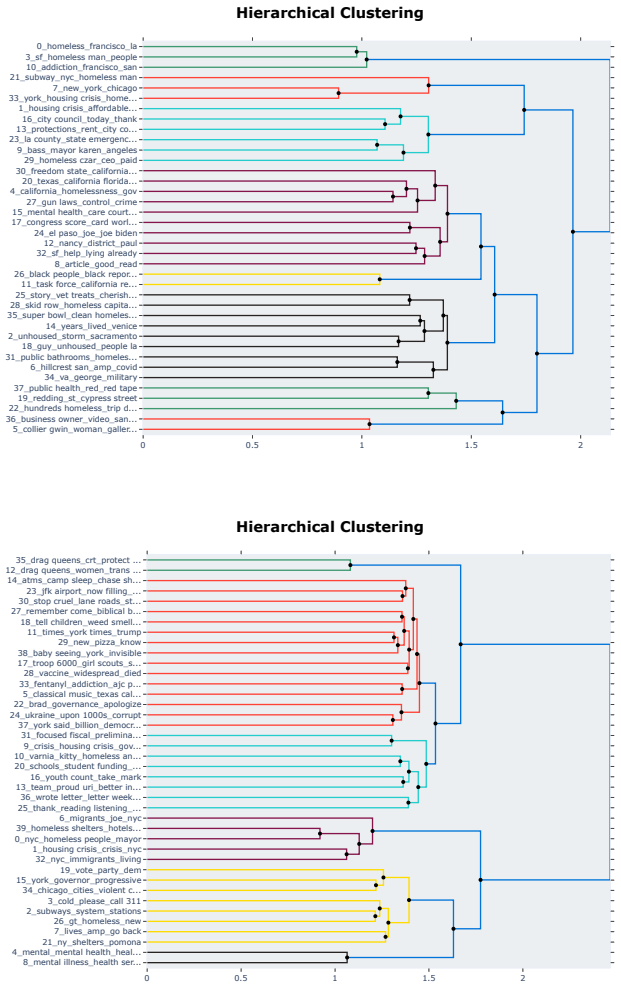


Fig. 15: Hierchical clustering of topics: if a split occurs under the set threshold of 1.4, topics are grouped together with the same color. On top, is the plot of the California dataset, and on the bottom is the one of the New York dataset.

of. In both cases, however, the number of generated clusters is close to the number that was spotted in the previous plots, even though the meaning is different.

In Fig. 16, document embeddings are projected into a 2D plane where each tweet is colored according to the topic it is assigned to. For both datasets, there is a macro-topic that is scattered almost everywhere while only a couple of the other major topics/clusters are visible. We need to remember that we are trying to visualize very high-dimensional objects on a 2D plane. In order to have a clearer visualization, we repeat the plot excluding this macro-topic and only including the most frequent and relevant topics for which a meaningful title is assigned. The updated result is shown in Fig. 17. It is now easier to spot the different clusters: some are more compact like the "black reparations" and "gallery owner arrested" while others like "drug addiction" are more spread around. This means that the formers are more specific and self-contained

Fig. 16: 2D visualization of tweets with the corresponding topic color coded. On top, is the plot of the California dataset, and on the bottom is the one of the New York dataset.

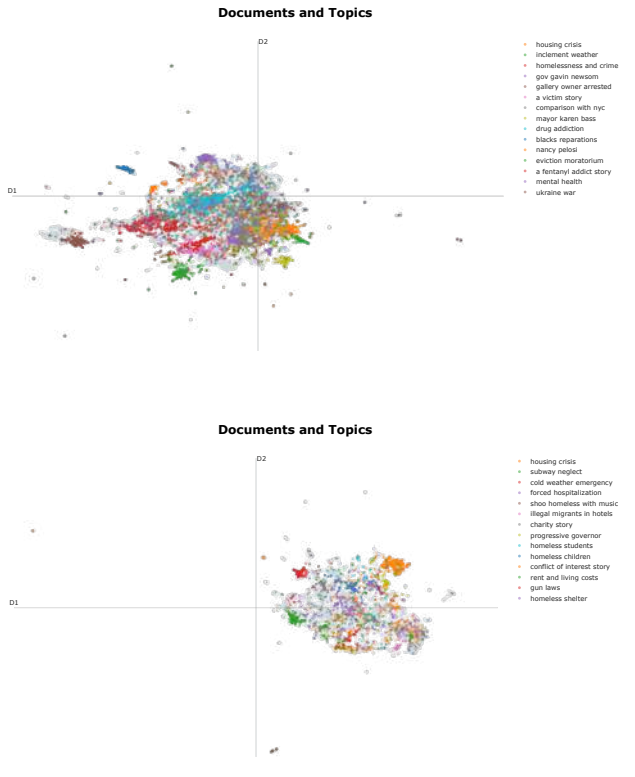


Fig. 17: 2D visualization of tweets with the corresponding topic color coded. Only the most frequent and relevant topics are present: the largest macro-topic is also hidden for clarity. On top, is the plot of the California dataset, and on the bottom is the one of the New York dataset.

topics while the latter is a broader topic with multiple points of view. Overall, both datasets show that tweets clusters are close to one another and they only span a small region. The only exception is the ”shoo homeless with music” topic of the New York dataset, which is far apart from the rest.

In order to analyze and truly understand the meaning of the topics extracted with BERTopic, one has to go through some of the most representative tweets for each. In Tab. 9 we report the number of tweets assigned to each of the most frequent topics, assign them a meaningful title, and include an exemplary tweet while providing a brief explanation. The same is done for the New York dataset in Tab. 10.

As noticed beforehand, for both datasets there is a macro-topic on homelessness to which a large portion of tweets is assigned. This is a very general theme that contains some aspects that are then deepened in the other topics. Overall, there are both common and shared topics and others that are specific to only one of the two states. Examples of the former are the housing crisis and the cold weather. The lack of affordable housing and the prohibitively high rents and living costs are one of the root causes of the homelessness crisis and are affecting not only the unemployed or the poorest sections of the population but also families, children, and students. The cold weather is an additional life-threatening struggle for those that are unhoused: while this is typical and expected in New York, the unusual floods that hit California are unprecedented and a consequence of climate change. Mental health is another subject present in both datasets but in different flavors. While in California the focus is on the fact that a large portion of the homeless population suffers from mental health conditions, in New York, there is a lot of debate around a law that allows the forced hospitalization of homeless that are affected by any mental condition. Other major topics present in the California datasets are the relationship between crime and drug abuse with homelessness: drugs are very common among the homeless as a coping mechanism. This also leads to committing assaults to get money while at the same time, unhoused people are more vulnerable to violent acts themselves. There are also lots of references to the situation in New York, thus validating our choice for a comparison. On the other hand, in the New York dataset, the focus is on immigrants that are hosted in hotels while American citizens and war veterans are left on their own living in the streets. At the same time, people are complaining about the neglect and filth of the subway system, with huge repercussions on the sanitary and hygienic conditions for those who use it as a shelter. Another common theme is the criticism of political figures: some examples are the mayor of Los Angeles Karen

id	Count	Keywords	Title	Representative tweet	Explanation
0	12710	homeless, francisco, la, state, city, gallery owner, crime, count, live, help	california's homeless	San Francisco has one of the biggest and highest homeless and poverty rates in the USA today.	Macro-topic on homelessness in California's major cities.
1	1451	housing crisis, affordable housing, building, people, homelessness crisis, city, solve, sf, income, need	housing crisis	Affordable housing should be equally addressed, outrageous rent prices in Los Angeles and (SF) is a major part of the problem. With so many coming to LA to be homeless in a warmer climate, those born/worked and paid taxes in LA should be served first	High rents and lack of affordable housing are among the leading cause of the homelessness crisis.
10	1069	addiction, francisco, san, california, amp, don, city, sf homeless, free, homeless addicts	drug addiction	if you BAN public drug usage in SF...less "homeless" drug addicts from other states and cities will not be interested in going to SF. Your homeless situation would be manageable and for people who are legitimately going through hard times, not mentally ill and drug addicts	Drug addiction and abuse are common among homeless people.
3	1044	sf, homeless man, people, white, homeless violence, city, even, video, arrest, likely	homelessness and crime	The data fom San Diego is clear. Homeless people are MUCH more likely to be involved in crime than the rest of the population. Homeless people are 130 TIMES MORE LIKELY to commit assault than the general population	Relation between homelessness and crime rates: both committed and suffered by homeless people.
7	841	new, york, chicago, eric adams, homeless nyc, mayor, asylum, nypd, cities, la	comparison with nyc	In a few years New York City will become just like LA and San Francisco in terms of homelessness. Great liberal policies are catching up.	Comparison between the homelessness crisis happening in California's major cities and New York: reference to the forced hospitalization implemented by NY's major.
2	716	unhoused, storm, sacramento, climate, beds, shelter program, inclement weather, homeless people, weather shelter, flooded	inclement weather	As California is hit hard by storms, few are facing as stark a challenge as the state's more than 170,000 homeless residents.	The unusual cold and the floods are additional challenges that the homeless have to face.
4	601	california, homelessness, gov, taxes, california governor, gov gavin, newsom wants, governor newsom, budget, will	gov gavin newsom	Gavin Newsom has destroyed California. What once was a beautiful place now is taxed to death and ravaged with homelessness. Stupidity is growing and we have to stop clowns like this from ruining our country.	Criticism directed at the governor of California Gavin Newsom for his policies addressing the homeless problem.
6	542	hillcrest san, amp, covid, experiencing homelessness, seniors, diego hospital, hi, hotel rooms, care, nurse speaks	a victim story	RIP Mr B discharged from a hospital in Hillcrest San Diego to the streets with no where to go as an older adult Black male experiencing homelessness.	A local story about a homeless black man who died after being discharged from a hospital.
9	514	bass, mayor karen, angeles, mayor bass, angeles mayor, venice, homeless, new mayor, 000 people, plan	mayor karen bass	Mayor Karen Bass of LA talks homelessness while two hours south our fed govt lets in over 8,000 illegals per day. Wise policy, Dems.	Criticism directed at the mayor of Los Angeles Karen Bass for her policies addressing the homeless problem.
15	497	mental health, care court, people, treatment, california, civil rights, health homelessness, senate, new health, chair prioritize	mental health	In California, a large proportion of chronically homeless individuals suffer from mental health conditions. The upcoming event on February 1 will discuss the relationship between homelessness and the health care system.	A large portion of homeless people suffer from mental health conditions: the healthcare system is asked to address this situation.
5	404	collier gwin, woman, gallery owner, art, francisco, hose, hosing, gwin gallery, spraying homeless, warrant	gallery owner arrested	Collier Gwin, the gallery owner who was filmed spraying a homeless woman with a hose in San Francisco, has been arrested for battery.	A local news story: Collier Gwin, an art gallery owner, has been arrested after spraying a homeless woman with a hose.
12	310	nancy, district, paul, san francisco, pelosi district, maga, house, homelessness, pelosi san, needles	nanci pelosi	Nancy Pelosi's district in San Francisco: the streets are covered in human feces, dirty needles, homeless camps…	Criticism towards Nancy Pelosi, since she is the congress representative of the district that includes San Francisco.
14	307	years, lived, venice, homeless, daughter fentanyl, failed save, says, jessica, car, later daughter	a fentanyl addict story	She failed to save her daughter from fentanyl's grip. A year later, her daughter is still homeless in San Francisco - and the streets have not changed.	A local story about a mother who couldn't save her daughter from fentanyl addiction who now is homeless.
17	307	congress score, card world, highest incarceration, sending, medical bankruptcy, america, 2022, billions dollars, world, homeless	ukraine war	Let's just keep sending billions to the Ukraine every month. Forget about the border, homeless people, fentanyl overdoses, veterans, American citizens or California. Ukraine is more important than America.	Questions and criticism about sending money to help Ukraine in the war against Russia with all the problems that plague America, among which the homelessness crisis.
13	301	protections, rent, city council, tenant protections, los, eviction moratorium, will, covid 19, expire, homelessness	eviction moratorium	Los Angeles renters are set to get new rights against eviction and large rent hikes just as the city's COVID protections are about to expire. Get the details from today's city council meeting.	Concern about the expiring covid-19 moratorium on evictions: there is a concrete risk of more people becoming homeless.
11	293	task force, california reparations, reparations task, blacks owed, black homeless, demands black, slavery, 1m homeless, make demands, give	blacks reparations	California would rather give $5M in reparations to people who were never slaves, than take care of their 69,000 homeless. WTH is wrong with these numbskulls????	There is a task force that is evaluating the possibility to give money to black citizens as reparation for past slavery. Some critics this pointing out the homeless crisis.

TABLE 9: Most relevant and frequent topics extracted from the California dataset. For each one, the keywords, the assigned title, a representative tweet, and a brief explanation are reported.

id	Count	Keywords	Title	Representative tweet	Explanation
0	3790	nyc, homeless people, mayor, shelter, streets, hotels, adams, will, many, help	homeless in nyc	New York City has 67,150 homeless people, including 21,089 homeless children, sleeping each night in New York City's main municipal shelter system. Yet somehow, they were able to house migrants at luxury hotels for months. Dems support illegals OVER America	Macro-topic on homelessness in New York City.
1	735	housing crisis, crisis, nyc, affordable housing, apartments, homeless, compact, nyc housing, rise, solve	housing crisis	The Mayor also reaffirmed the City's commitment to addressing NYC's housing shortage and solve its affordable housing crisis.	The accommodation shortage and the lack of affordable housing are one of the root causes of the homeless crisis in NYC.
39	251	homeless shelters, hotels, open 24, unhoused, public, need, toilet, ny, building, shelter cont	homeless shelter	NYC provides comfy cots, 3 meals a day, sanitary facilities, warm or AC facilities for allowed by , yet many homeless waste away after serving this under bridges and side streets.	The city provides shelter for the homeless, but most of them end up living on the streets anyway.
32	164	nyc, immigrants, living, rent, homeless, us, free, taxes, benefit workers, wealthiest	rent and living costs	Man I know about of folks in nyc like myself living from paycheck to paycheck! The housing prices/rent is ridiculously high. And has been more people that became homeless in the last two years than the previous decade. That's a fact	Even employed people risk becoming homeless due to the prohibitively high rents and living costs.
2	157	subways, system, stations, see, homeless people, manhattan, nyc train, public, clean, public transit	subway neglect	The NYC Subway is polluted, the air quality in many stations is really bad. The smell of trash, the homeless, on top of these pollutants. You have to wear something when riding the subway.	The subway system in NYC is used as a shelter by many homeless: there are many concerns about the hygiene and degradation of this phenomenon.
15	135	york, governor, progressive, today, states, blah blah, washington, amp, california new, crime	progressive governor	Yup, progressive governing has worked GREAT in NYC. Thriving economy, low crime (so low), growing population,...	Criticism against the progressive policies of the mayor of NY and the governor of New York State, both belonging to the Democratic party.
3	125	cold, please, call 311, code blue, nyc, one, night, will denied, intake homeless, temperatures expected	cold weather emergency	Due to frigid temperatures expected tonight, NYC is activating , beginning at 4pm. No one will be denied intake at a homeless shelter. Please do not be indifferent to any vulnerable person you see in the street. Call 311 immediately.	Due to the extreme cold weather typical of NYC, an emergency number 311 is activated to provide assistance to all homeless.
21	125	ny, shelters, pomona, mobile units, will, worked, health services, income, homeless families, day	homeless children	Guess who lives in family shelters? Kids! Children, who go to "schools, day care centers, dance, tutoring and child care centers and parks"... Shouldn't they have the same opportunities as their peers? What about *their* safety?	There are families with children that are homeless and have to live in shelters while getting their education.
4	109	mental, mental health, health expert, questions forced, forced hospitalization, judge, york homeless, proceed, people, let deaf	forced hospitalization	A judge ruled that New York City's controversial plan to allow first responders to involuntarily hospitalize homeless people with mental illnesses can proceed.	A controversial law in NYC allows the forced hospitalization of homeless people that suffers from mental illness.
6	106	migrants, joe, nyc, vets homeless, hotels, president, homeless biden, help, shelters, 500 night	illegal migrants in hotels	President Biden didn't bother visit the border crisis in Midtown. It is outrageous that illegal migrants refuse to move to a different shelter; demand hotel accommodations. Not only can NY not afford it, but we have homeless NYers on our streets.	Protests against President Biden for the fact that illegal migrants are hosted in luxury hotels while American citizens are on the streets.
7	85	lives, amp, go back, homeless, charity providing, prison now, cynthia, vanessa santiago, queens runs, vee	charity story	Vanessa Santiago served 17.5 years in prison. Now she lives in Queens, runs a small charity providing free furniture and more to people in need	Local story of an ex-convict woman that now helps people in need with free furniture.
34	83	chicago, cities, violent crime, gun laws, shootings, address, liberal, control, every, state	gun laws	Which nations specifically? Also make sure to add in contributing factors that involve violent crime. Including poverty, homelessness, and cultural standards. How come jurisdictions with the strictest gun regulations in the United States (Chicago, New York) have more gun crimes?	Relation between gun laws and violent crime including the effect of homelessness on the latter.
20	74	schools, student funding, homeless students, fair, formula, nyc, high need, libraries, budget, proposed changes	homeless students	NYC proposes new funding stream for homeless students and high-need schools	The homelessness problem also affects students: new founding are being proposed to address this issue.
22	74	brad, governance, apologize, canada, will, tor fire, tory gaslighting, rely natural, liberal ll, refuses job	conflict of interest story	Get a load of the big hypocrisy by Brad, this guy is making it impossible for citizens to see the conflicts of interest of his wife's homeless service provider clients that are in charge of the city's dangerous homeless shelters.	A local story about the possible conflict of interest between the city comptroller Brad Lander and his wife's homeless service provider clients.
5	64	classical music, texas california, use classical, york use, music shoo, shoo homeless, people, opera, eleven convenience, using roaring	shoo homeless with music	7-Eleven stores in Texas, California, New York use classical music to shoo homeless people.	Some 7-Eleven stores are blasting classical music to keep homeless people away.

TABLE 10: Most relevant and frequent topics extracted from the New York dataset. For each one, the keywords, the assigned title, a representative tweet, and a brief explanation are reported.

Bass, the governor of California Gavin Newson, and the San Francisco congress Representative Nancy Pelosi. The same happens for the mayor and governor of New York. All these political figures are attacked for their policies and for being part of the Democratic party. Local stories lead an important part of the discussion: an example is the arrest of an art gallery owner in San Francisco who spray a homeless woman with a hose. This news sparked a debate with supporters on both sides, in favor or against the arrested. In the New York dataset, a similar story is the one about 7-Eleven stores that play high-volume classical music to keep homeless people away. Finally, other topics that appear with lower frequencies are the Ukraine war, people questioning the money sent to help Ukrainians instead of the homeless, and the covid-19, with fear of the consequences that the end of a moratorium on evictions issued during the pandemic could bring.

V. Sentiment Analysis

While topic modeling gives much insight into what people are talking about, they can only partially describe the overall people's sentiment towards the homeless crisis. To achieve this we adopt a `Twitter-RoBERTa-base for Sentiment Analysis` model, that is publicly available [20]. This is an updated version [21] of a RoBERTa-base model trained on ∼124M tweets from January 2018 to December 2021, and finetuned for sentiment analysis with the TweetEval benchmark [22]. RoBERTa (Robustly optimized BERT approach) is a transformer-based language model that is a modified version of the BERT (Bidirectional Encoder Representations from Transformers) model, which is trained on a large corpus of text data in an unsupervised manner. The version used in this work assigns to each tweet the probability of belonging to one of the following 3 classes: "Negative", "Positive" or "Neutral". The class with the highest probability becomes the label of that tweet.

The preprocessing step is the same as described in Section IV. Again, leveraging GPU acceleration the model classifies the tweets from both datasets pretty quickly.

In Fig. 18 we showcase the overall distribution of the three classes for both datasets. The two results are pretty close to each other: the majority of tweets are classified as "Negative", only 7-8% as "Positive", and the rest, around a third, as "Neutral".

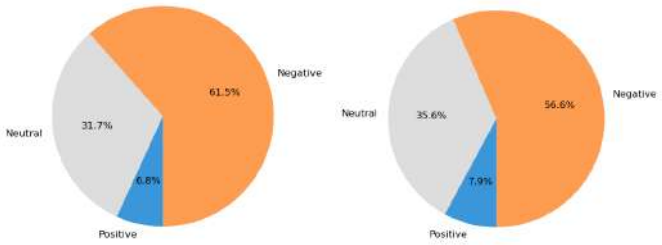


Fig. 18: Distribution of negative, positive, and neutral classified tweets of California (left) and New York (right) datasets.

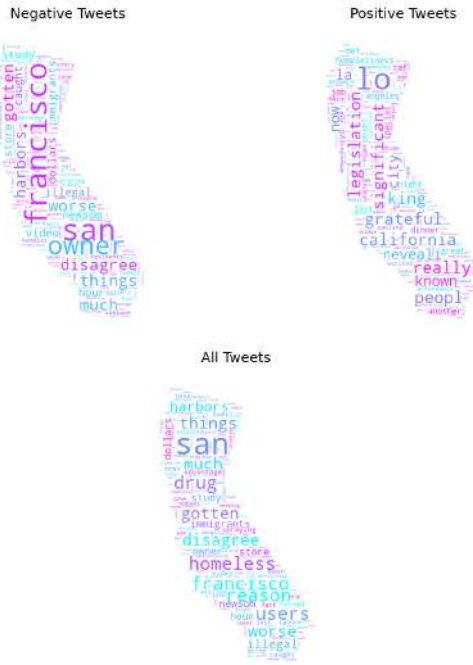


Fig. 19: Word clouds for negative (top-left) and positive (top-right) classified tweets of the California dataset. On the bottom, there is the one for the complete dataset.

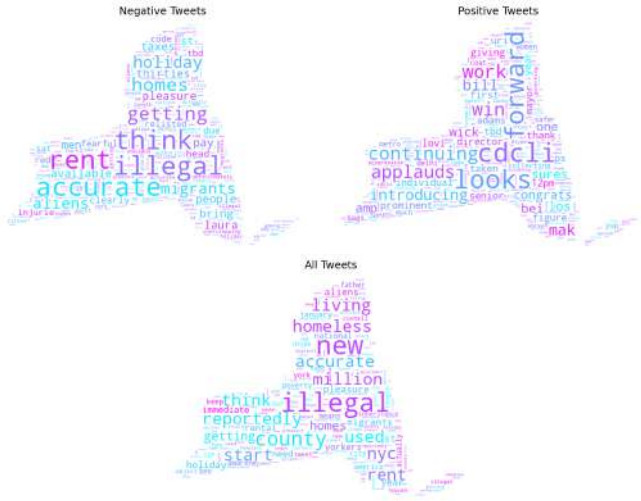


Fig. 20: Word clouds for negative (top-left) and positive (top-right) classified tweets of the New York dataset. On the bottom, there is the one for the complete dataset.

To understand the meaning behind this classification, similarly to what we performed in Section IV, we report some significant tweets for each of the three labels in Tab. 11 and in Tab. 12 for the California and New York dataset respectively. Firstly, tweets labeled as "Neutral" are simply news titles or story headlines which are detached and unemotional by default. "Negative" tweets include criticism, insults, swear words and slurs and are mainly addressed to politicians or to express disgust with the neglect and all the homeless-

Classification	Count	Representative tweets
Negative	16080	Fuck your issuing an arrest warrant for this man for doing your job? Fuck that entire shit city. Fix your streets you shitty city with a great football team.
		Not only that, California is in the worst state it has been in my lifetime. The homeless problem is out of control, its infrastructure is crumbling and the schools are failing.
		Politicians SUCK doing nothing for homeless they put on the street punish business owner for trying to run a business. FUCK California!! probably why everyone is moving out of state?????
		You refuse to clean up your state! It's riddled with homeless junkies who spread disease and violence! California is the asscrack of america
Positive	1783	I am excited to share that we received a $10,000 donation last night to the New Beginnings Scholarship Fund for Unhoused Youth in San Diego!
		Thanks and gratitude to our federal partners for providing $75M to help unsheltered Californians!
		Thank you to over 70 volunteers who reported for duty at yesterday's Greater Los Angeles Homeless Count! It was wonderful
		The Encino and Los Angeles teams joined forces with Our Big Kitchen Los Angeles to make a difference. They managed to create an incredible 205 meals and 205 cookies for people in need!
Neutral	8273	Kymberli and B first lived in their car. They then lived in a vacant lot. Kymberli and B now live in a tent community in Oakland, California.
		San Jose officials plan to send workers to communicate evacuation orders to unhoused residents along the Guadalupe River
		The 2023 Greater Los Angeles Homeless Count, a point-in-time snapshot of in County that helps determine the distribution of and services to the unhoused, began Tuesday night.
		Photos: Volunteers Survey San Diego's Homeless Population San Diego Search Engine Optimization

TABLE 11: Summary of sentiment classification for the California dataset: for each class four representative tweets are reported.

Classification	Count	Representative tweets
Negative	4035	"Senator, why you send all my $ to Ukraine when people are dying in Harlem? People are homeless in the Bronx! 50 yrs in Harlem & still F'ed up! You are useless!"
		F$#& these migrants bitches help the homeless we have in NYC. What is wrong with the government.
		Nothing good in NY it's a dirty dumpster, the football team are a joke. Full of homeless, hookers, criminals and sewer rats.. overpriced dirty apartments lmao
		Manhattan is dirty, stinks of marijuana, homeless and crime everywhere!! City has been ruined in last few years.
Positive	562	Also I recommend on Queens Square, a wonderful cafe that do wonderful things for the community and homelessness and they do wonderful red velvet cake so get down there
		So proud of the work we do at the Y! Bridgeport homeless shelters prepare for frigid weekend.
		This are fantastic news. We need solutions to the NYC affordable housing crisis.
		The Sleeping Bag Project NYC. DR. JODI has a great grassroots organization giving out sleeping bags to the homeless. A $25 donation goes a long way. He has no expenses at all Happy Birthday
Neutral	2538	Today at 4 p.m. NYC's Department of Homeless Services Code Blue goes into effect. It will last until 8 a.m. tomorrow morning.
		New York City to Open Migrant Shelter at Brooklyn Cruise Terminal
		"New York City Sparks Debate by Using New York's Mental Hygiene Law to Engage the Homeless Population"
		NYC's Involuntary Hospitalization Plan Can Proceed, A Judge Rules. The decision comes a few months after the controversial plan was introduced by Mayor Eric Adams.

TABLE 12: Summary of sentiment classification for the New York dataset: for each class four representative tweets are reported.

related problems. Finally, "Positive" tweets are essentially about helpful initiatives such as monetary and merchandise donations, and volunteer actions, and they include information about shelter and weather conditions.

In Fig. 19 we plot three word clouds, taking care of excluding stop words, shaped as the California state: one for tweets classified as "Negative", one for "Positive", and one for the entire collection. We use the latter as a baseline to compare which words emerge as more frequent for the two classes. The same is done in Fig. 20 for the New York dataset using the corresponding shape. California's negative tweets show the predominance, excluding words related to locations, of terms like "worse", "owner" and "gallery" which refer to the local story described in Section IV, "dollars", "things" and "stores" etc.... This highlights the fact that people are most upset about the damages and financial losses that homeless people might cause to businesses and how taxpayer money is spent to address the problem. On the other hand, positive tweets are represented by words like "grateful", "significant", "legislation", "people", "dinner", "special" etc.... These words reinforce the concepts of donations and volunteer actions while appealing to changes in the legislation to help solve the homelessness crisis. In the overall word cloud, one can spot "drug" and "users" among other terms. These refer to the drug addiction problem, a theme present regardless of the classification of the tweets. Moving to the New York dataset, negative tweets are described mainly by words like "rent", "homes", "illegal", "migrants" and "aliens". In this case, the focus is twofold: on one hand, people criticize the high rents and the lack of affordable housing,

on the other, they blame immigrants for taking resources and rooms in hotels that could be used by American citizens. Similar to Califonia's case, positive tweets express gratitude for community initiatives with words such as "applauds", "looks", "forward", "win", "work", and "continuing". The acronym "cdcli" is also present: it stands for Community Development Corporation of Long Island, which is a non-profit organization whose mission is to address the growing demand for affordable housing. Finally, in the overall word cloud, no specific term appears that is not present in the previous ones.

VI. Conclusions

By combining network science tools and topic modeling we have gathered some key insights into the complex problem of homelessness. In particular, we successfully identify some of the causes: increase in rents, and lack of affordable housing which affects not only poor and unemployed people but also students and low-income families, children included. We then uncovered the relationship between drug abuse, mental illness, and homelessness while understanding which are the everyday struggles that homeless people have to deal with, such as the cold and inclement weather.

The comparison between California and New York was useful to highlight which are the essential, location-independent themes, like the aforementioned ones, distinguishing others that are specific to each state, such as crime in California and immigration and the neglect of New York's subway. On the other hand, the sentiment analysis shed light on people's opinions: the vast majority show negativity towards the homeless. especially if immigrants, and political figures, democrats in both states, are considered responsible for the decay connected with homelessness. The minority, instead, focuses on sharing helpful initiatives that aim to alleviate the hardships of the homeless life while advocating for more decisive policies to tackle the problem in a systematic way.

While we can be satisfied with the obtained results, one can always improve and investigate further by collecting a larger number of tweets, particularly in different periods of the year to also perform a temporal analysis. It would also be interesting to perform more advanced sentiment analysis to classify tweets more specifically, for example distinguishing between fear, hate, joy etc.... At this stage, we only separate positive and negative ones, highlighting the fact that the latter are the vast majority.

Finally, this project has been a great opportunity to implement and utilize various techniques from both the network science and natural language processing fields. Both R and Python were used, alongside Gephi for the visualization of the networks, to run the algorithms and the machine learning models adopted for this work.

References

[1] U. D. of Housing and U. Development, "The 2021 annual homeless assessment report (ahar) to congress: Part 1," 2021.

[2] N. A. to End Homelessness, "Causes of homelessness," 2021.

[3] C. B. . P. Center, "California's high housing costs: Causes and consequences," 2021.

[4] C. D. of Housing and C. Development, "Homelessness in california: State of the state," 2021.

[5] "Twitter API." https://developer.twitter.com/en/docs/twitter-api. [Online; accessed 13-March-2023].

[6] "Google Trends." https://trends.google.com/home. [Online; accessed 13-March-2023].

[7] "Natural Language Toolkit." https://www.nltk.org/#natural-language-toolkit. [Online; accessed 13-March-2023].

[8] "Gephi software." https://gephi.org/. [Online; accessed 13-March-2023].

[9] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," *PLOS ONE*, vol. 9, pp. 1–12, 06 2014.

[10] A.-L. Barabasi and M. Posfai, *Network science*. Cambridge: Cambridge University Press, 2016.

[11] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: A python package for analysis of heavy-tailed distributions," *PLoS ONE*, vol. 9, p. e85777, jan 2014.

[12] "Pearson's r implementation." https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.assortativity.degree_pearson_correlation_coefficient.html#networkx.algorithms.assortativity.degree_pearson_correlation_coefficient. [Online; accessed 13-March-2023].

[13] "Double edge swap implementation." https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.swap.double_edge_swap.html. [Online; accessed 13-March-2023].

[14] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[15] "Sentence transformer model." https://huggingface.co/microsoft/mpnet-base. [Online; accessed 13-March-2023].

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

[17] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[18] D. D. Xu and S. B. Wu, "An improved tfidf algorithm in text classification," in *Material Science, Civil Engineering and Architecture Science, Mechanical Engineering and Manufacturing Technology II*, vol. 651 of *Applied Mechanics and Materials*, pp. 2258–2261, Trans Tech Publications Ltd, 11 2014.

[19] J. Carbonell and J. Stewart, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 06 1999.

[20] "Twitter-RoBERTa-base for sentiment analysis." https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. [Online; accessed 13-March-2023].

[21] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," 2022.

[22] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Proceedings of Findings of EMNLP*, 2020.