Master's Program in Business Analytics and Data Science

Department of Business Administration

<u>Master's Dissertation</u>

Analyzing the Refugee Crisis in Greece through Twitter: topic modelling and network analysis

Ανάλυση της Προσφυγικής Κρίσης στην Ελλάδα μέσω του Twitter: θεματική μοντελοποίηση και ανάλυση δικτύων

Spyridoula (Leela) Papadioti

Supervising Professor: Dr Andreas Georgiou

Submitted as a requirement for the MSc in Business Analytics and Data Science

February 2023

Abstract

Social media platforms have become an integral part of the public discourse and researchers continue to turn to data mined from social media to understand public opinions on a variety of topics. This study investigates the public discussion on Twitter regarding the refugee crisis in Greece throughout the period 2015 – 2021, using a total of 116,530 mined tweets utilizing the hashtag #RefugeesGR. This hashtag proved crucial for mining data that captures a wide spectrum of discussion around refugees, as it was extensively used by a large variety of Twitter account types, including verified, professional and personal accounts. The dataset of mined tweets was analyzed via a topic modelling algorithm based on Latent Dirichlet Allocation and yielded results that were on par with the analysis of real-world events from that period in Greece. Topics that maintained popularity throughout the years include refugee boat arrivals on the Greek islands, the living conditions in formal and informal refugee camps across Greece and calls-to-action for humanitarian aid and in solidarity with refugees, while a shift was observed in the geographic-related discussion on refugees from the border region of Greece with North Macedonia during 2015 – 2016 to the border region with Turkey from 2017 onwards. The interactions, particularly mentions, between Twitter users using the hashtag #RefugeesGR were investigated using network analysis and visualization, revealing that the most influential accounts within the network belonged to International Humanitarian Organizations, EU governmental bodies, and individual (international and local) politicians. Social network analysis was enhanced with community detection, forming clusters of accounts that were not grouped together based on similar characteristics in terms of user type, but rather on the accounts which were more frequently called-to-action at the same time, thus bringing together accounts that were co-relevant within a specific time-period. This study serves as an attestation to the value of data mined from Twitter and the reliable insights on real-world events that could be derived from tweets.

Keywords: #RefugeesGR, refugees, refugee crisis, Greece, twitter, data mining, data scraping, twitter mining, twitter scraping, text cleaning, text pre-processing, natural language processing, NLP, topic analysis, topic detection, topic modelling, LDA, LDAvis, community detection, graphs, networks, network analysis, network metrics, network visualization, Gephi, social network, social network analysis, SNA, OSINT.

Περίληψη

Οι πλατφόρμες των μέσων κοινωνικής δικτύωσης έχουν γίνει αναπόσπαστο μέρος του δημόσιου διαλόγου και οι ερευνητές συνεχίζουν να χρησιμοποιούν δεδομένα που εξορύσσονται από τα μέσα κοινωνικής δικτύωσης για να κατανοήσουν τις απόψεις του κοινού για μια πληθώρα θεματικών. Αυτή η μελέτη ερευνά τη δημόσια συζήτηση στο Twitter σχετικά με την προσφυγική κρίση στην Ελλάδα κατά την περίοδο 2015 – 2021, χρησιμοποιώντας ένα σύνολο από 116,530 εξορυγμένα tweets που περιλαμβάνουν το hashtag #RefugeesGR. Το hashtag αυτό αποδείχθηκε κρίσιμο στην εξόρυξη δεδομένων που καλύπτουν ένα ευρύ φάσμα συζητήσεων γύρω από το προσφυγικό, καθώς χρησιμοποιήθηκε εκτενώς από διαφόρων τύπων λογαριασμών Twitter, συμπεριλαμβανομένων επαληθευμένων, επαγγελματικών και προσωπικών λογαριασμών. Η βάση δεδομένων των tweet που εξορύχθηκαν αναλύθηκε μέσω ενός αλγόριθμου θεματικής μοντελοποίησης που βασίζεται στη Latent Dirichlet Allocation και τα αποτελέσματα αυτής συνάδουν με την ανάλυση αληθινών γεγονότων εκείνης της περιόδου στην Ελλάδα. Οι θεματικές που διατήρησαν υψηλή δημοτικότητα ανά τις χρονιές περιλαμβάνουν τις αφίξεις βαρκών με πρόσφυγες στα ελληνικά νησιά, τις συνθήκες διαβίωσης σε επίσημες δομές φιλοξενίας και ανεπίσημους προσφυγικούς καταυλισμούς της Ελλάδας, καθώς και τις εκκλήσεις σε δράση για παροχή ανθρωπιστικής βοήθειας και προς αλληλεγγύη με τους πρόσφυγες, ενώ παρατηρήθηκε μια στροφή στη γεωγραφική συζήτηση του προσφυγικού από τη συνοριακή περιοχή της Ελλάδας με τη Βόρεια Μακεδονία κατά την περίοδο 2015 – 2016 στη συνοριακή περιοχή με την Τουρκία από το 2017 και έπειτα. Οι αλληλεπιδράσεις μεταξύ χρηστών του Twitter που κάνουν αναφορά στο hashtag #RefugeesGR διερευνήθηκαν χρησιμοποιώντας ανάλυση και οπτικοποίηση δικτύων, αποκαλύπτοντας ότι οι λογαριασμοί με τη μεγαλύτερη επιρροή στο δίκτυο ανήκουν σε Διεθνείς Ανθρωπιστικούς Οργανισμούς, κυβερνητικούς φορείς της ΕΕ, καθώς και σε προσωπικούς λογαριασμούς (διεθνών και τοπικών) πολιτικών προσώπων. Η ανάλυση του κοινωνικού δικτύου ενισχύθηκε μέσω της ανίχνευσης κοινοτήτων, σχηματίζοντας συστάδες λογαριασμών που δεν ομαδοποιήθηκαν με βάση τα παρόμοια χαρακτηριστικά τους όσον αφορά στον τύπο χρήστη, αλλά με βάση τους λογαριασμούς που καλούνταν πιο συχνά σε δράση ταυτόχρονα, ομαδοποιώντας έτσι μαζί λογαριασμούς που ήταν ταυτοχρόνως σχετικοί σε μια συγκεκριμένη χρονική περίοδο. Αυτή η μελέτη αποτελεί μια επιβεβαίωση ως προς την αξία των δεδομένων που εξορύσσονται από το Twitter, καθώς και των αξιόπιστων πληροφοριών σχετικά με γεγονότα του πραγματικού κόσμου που θα μπορούσαν να προκύψουν από τα tweets.

Preface

This body of work was inspired by all the brave people I have met, the stories I had the privilege to listen to and the response I am proud to have been a part of during the 2015 – 2021 "refugee crisis" era in Greece.

My motivation for this dissertation comes from the realization of the gravity of the refugee movements in the post-2015 period on the ground in first reception European countries, like Greece, and the major impact this fact has brought along in such countries, including the need for the establishment of adequate response mechanisms, social protection floors and legal systems to ensure that people on the move, along with local populations, have de facto access to human and refugee rights. Additionally, the rise of the far-right movements across European countries as a result of a discriminatory and prejudiced rhetoric that is –in many cases– conflicting with the sincerely hospitable nature of their people, was always puzzling for me to witness and understand. Especially when, in the very same countries, the solidarity efforts and vast networks of regular citizens that joined forces and volunteered their time and resources in order to provide all-around material and social support to refugees coming in their cities led to the establishment of important civil society actors that have supported not only people on the move and refugees, but filled important gaps for local society as well, coming in direct contrast with trending political agendas.

Working in the humanitarian sector as an information management and data analytics professional throughout this period in Greece and having access to official data and figures of the situation, I have a multifaceted understanding of the reality on-the-ground. However, through the work in this dissertation, I chose to explore more the side of the refugee crisis having to do with the stories that get to the public, the chit-chat around the news and what people actually talk about on the subject. Since social media is such an irrefutable part of our everyday lives and a regular means of expressing our thoughts, Twitter seemed like the ideal place to focus my investigation.

I would very much like to express my gratitude towards my supervising professors for their continuous support and encouragement to choose a topic so important to me, as well as my family, friends and partner for their understanding and patience.

Home[1]

no one leaves home unless
home is the mouth of a shark
you only run for the border
when you see the whole city running as well

your neighbors running faster than you
breath bloody in their throats
the boy you went to school with
who kissed you dizzy behind the old tin factory
is holding a gun bigger than his body
you only leave home
when home won't let you stay.

no one leaves home unless home chases you
fire under feet
hot blood in your belly
it's not something you ever thought of doing
until the blade burnt threats into
your neck
and even then you carried the anthem under
your breath
only tearing up your passport in an airport toilet
sobbing as each mouthful of paper
made it clear that you wouldn't be going back.

you have to understand,
that **no one puts their children in a boat**
unless the water is safer than the land
no one burns their palms
under trains
beneath carriages
no one spends days and nights in the stomach of a
truck
feeding on newspaper unless the miles travelled
means something more than journey.
no one crawls under fences
no one wants to be beaten
pitied

no one chooses refugee camps
or strip searches where your
body is left aching
or prison,
because prison is safer
than a city of fire
and one prison guard
in the night
is better than a truckload
of men who look like your father
no one could take it
no one could stomach it
no one skin would be tough enough

the
go home blacks
refugees
dirty immigrants
asylum seekers
sucking our country dry
niggers with their hands out
they smell strange
savage
messed up their country and now they want
to mess ours up
how do the words
the dirty looks
roll off your backs
maybe because the blow is softer
than a limb torn off

or the words are more tender
than fourteen men between
your legs
or the insults are easier
to swallow
than rubble
than bone
than your child body
in pieces.
i want to go home,
but home is the mouth of a shark
home is the barrel of the gun
and no one would leave home
unless home chased you to the shore
unless home told you
to quicken your legs
leave your clothes behind
crawl through the desert
wade through the oceans
drown
save
be hunger
beg
forget pride
your survival is more important

no one leaves home until home is a sweaty voice
in your ear
saying-
leave,
run away from me now
I don't know what I've become
but I know that anywhere
is safer than here

Warsan Shire

[1] www.youtube.com/watch?v=nI9D92Xiygo

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Societal relevance

According to the United Nations High Commissioner for Refugees (UNHCR), by the end of 2021 there were 89.3 million[2] forcibly displaced persons globally and with the recent war in Ukraine that number has reached above 100 million[3] during 2022, a stark figure meaning that 1 in every 78 people on earth has been forced to flee due to persecution, conflict, violence and human rights violations. To give these numbers a little context when it comes to the refugee situation in Europe, there are some 6.8 million refugees from Syria and 2.7 million refugees from Afghanistan worldwide, while the country with the largest refugee population is Turkey, hosting 3.8 million people.

Figure 1-1 below from UNHCR's operational data portal provides a visual representation of the pathways displaced persons take to reach Europe and the influx of asylum-seekers to first reception countries in 2022.

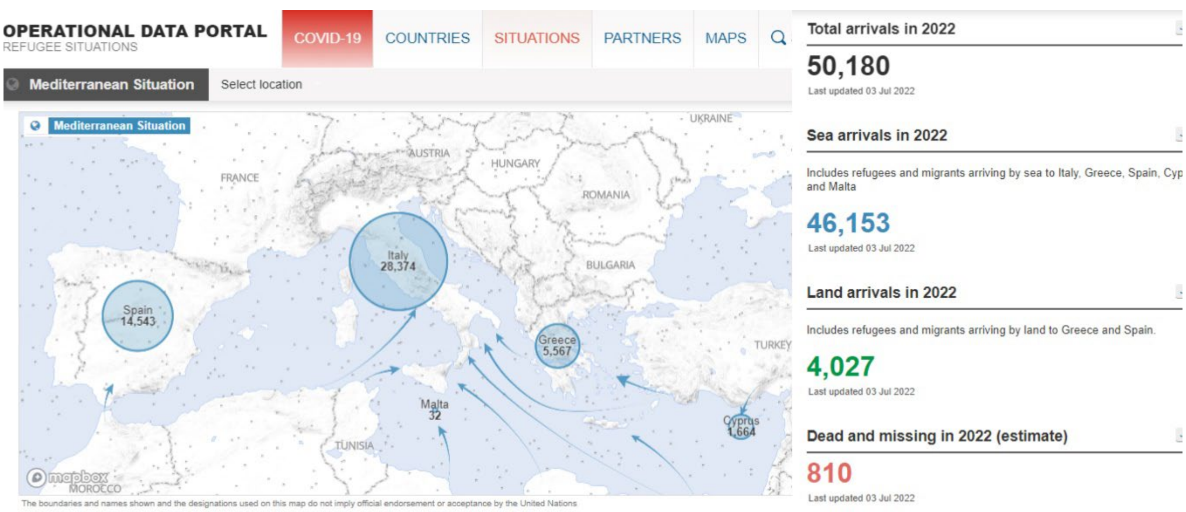

Figure 1-1: UNHCR Operational Data Portal Snapshot depicting arrivals pathways to Europe[4]

[2] UNHCR's 2021 Global trends report on forced displacement: www.unhcr.org/publications/brochures/62a9d1494/global-trends-report-2021.html

[3] https://www.unhcr.org/refugee-statistics/insights/explainers/100-million-forcibly-displaced.html#:~:text=As%20a%20result%20of%20these,the%20first%20time%20on%20record

[4] Snapshot as of 3 July 2022: data.unhcr.org/en/situations/mediterranean

Looking at Greece specifically, more than 1.2 million persons arrived in the country from 2015 to 2021, through both sea and land borders[5].

To better understand the high influx of refugees in Greece, mention needs to be made to the Syrian Civil War that began in 2011 and created a huge humanitarian crisis, which after eleven years still remains the world's largest refugee crisis to date. More than 6.6 million Syrians have been forced to flee their country and another 6.7 million people remain internally displaced. The Syrian Civil War dramatically increased the number of asylum seekers from 2014 onwards, particularly because of the harsh conditions in the neighboring countries of Syria, namely Jordan, Lebanon, and Turkey. Middle East states became increasingly unwilling to host refugees as the crisis intensified, leading to growing flows towards Europe. Most of these populations made their way to Greece via Turkey in numbers far greater than anything the country had experienced before.

During the period from January until July 2015 all European states had their borders open and, during that time, Frontex recorded a 663% increase[6] in border crossings compared to the same period in 2014. Although Syrians comprised the vast majority of asylum-seekers crossing the land and sea borders to Greece during 2015, there were also large numbers of Afghans and Iraqis among the people on the move.

This sudden change in the situation on-the-ground led to a humanitarian emergency in Greece, where UNHCR and multiple International Humanitarian NGOs drastically scaled-up their operations to assist the Greek government in its response. At the same time, the emergency situation brought international media attention to the region and the discussion on refugee movements became a popular topic in social media platforms.

This dissertation aims to utilize open-source intelligence (OSINT) to investigate the public opinion about refugees in Greece over the period 2015 – 2021 and specifically the public discourse on Twitter, which will act as a giant database of information for these purposes.

[5] data.unhcr.org/en/situations/mediterranean/location/5179
[6] frontex.europa.eu/assets/Publications/Risk_Analysis/WB_Q2_2015_report.pdf

1.2 Research goal

This dissertation will attempt to dive into the dialogue surrounding the refugee crisis in Greece during the period 2015 – 2021, using data mined from the Twitter platform, in an effort to comprehend the online discussion on refugees and its relevance to real-world events on the ground. The purpose of the analysis herein is to determine the main topics surrounding the discourse on refugees in Greece during the years 2015 – 2021 and observe how these topics shifted throughout the period at hand. Moreover, this dissertation will also focus on the Twitter users partaking in the discussion on refugees in Greece, including which are the most important and influential accounts, which communities are formed through a user interaction network and how these users can be classified within sectoral clusters. This work is based on the hypothesis that data mined from Twitter is suitable to provide reliable insight of the public discourse on real-world events and this potential can be harnessed through topic modelling for efficiency purposes.

1.3 Structure of the dissertation

The work presented in this dissertation is structured as follows. Chapter 2 provides an overview of some important and relevant pieces of research on why the social media platform Twitter has gained popularity among big data studies and how researchers have utilized the platform when investigating the refugee issue. Chapter 3 covers the methodology used to produce the analysis of Twitter data and in particular how the data was mined from Twitter and later cleaned on a step-by-step basis. Chapter 3 also presents a synopsis of the algorithm used for topic modelling (LDA), how the model can be evaluated and how the visualization can be presented. Furthermore, the field of social network analysis is outlined in Chapter 3, including the most important measures and node roles within a graph, but also how a network can be best visualized. Chapter 4 focuses on presenting the findings of the analysis via a series of visualizations on a year-by-year basis, aiming to a thorough understanding of the events that sparked the highest traffic on Twitter, as well as how these events translate to the use of particular keywords and hashtags. Chapter 4 also contains the findings of Topic Analysis through LDA and examines its output in contrast to the event analysis. Finally, the analysis and visualization of the network of Twitter users' interactions is presented in Chapter 4, offering an insight in the communities of users who drive the public discourse on Twitter regarding #RefugeesGR. This dissertation concludes with Chapter 5, summarizing the main findings of the work presented, as well as the limitations of the methodology followed and proposals for further exploration withing the same scope and focus.

2 Background and related work

There is an increasing trend in utilizing open-source intelligence in human migration studies from 2007 onwards, with more and more published studies undertaking this opportunity to use big data, due to the valuable knowledge that can be gained from analyzing information from online sources (Ashton et al., 2016).

A recent (2021) study tried to answer the question of "how big data can help to understand the migration phenomenon", by analyzing how different datasets and models can be used in quantifying and interpreting the three traditional phases of migration (journey, stay and return) (Sîrbu et al., 2021). The study found that in terms of estimating immigration flows and stocks, although there is existing research trying to use big data extracted from social media networks to nowcast immigration, it is lacking a proper methodological gold standard, as precise immigration rates on the present are generally unknown, and past rates can be noisy, leading to complications in the validation of nowcasting models. On the other hand, researchers believe that migrant integration can be measured using several new data types and even introduce novel integration indices, based on mobile data, social network language, sentiment and network analysis, however, the current relevant research is slightly less developed, mostly due to the low availability of ready-to-use datasets (Sîrbu et al., 2021).

A slightly older, but large study of 132 social media research papers between 1997 and 2017 (Kapoor et al., 2018) suggests that literature around social media has a very broad range and specifically revolves around the following thirteen themes:

1. Social media use, behaviors and consequences
2. Reviews and recommendations on social media
3. Social media and associated organizational impact
4. Social media for marketing
5. Participation in social media communities
6. Risks and concerns on the use of social media
7. Stigmatization of social media usage
8. Value creation through social media
9. Role of social media during critical/extreme events
10. Social media for help/support

11. Public bodies and social media interaction
12. Traditional vs social media
13. Testing pre-established models

This dissertation falls under two of the main themes above, namely the role that social media play in extreme events and also how social media is utilized to seek help and support. The work here will focus specifically on data mined from the Twitter social media platform, as research supports the argument that Twitter may be considered a reasonable source for information about immigration and refugee placement, while having the added benefit of the real-time dimension to the information being tracked (Aswad & Menezes, 2018).

2.1 Twitter and social network data

Twitter[7], launched in July 2006 by Jack Dorsey, is one of the most popular social networks worldwide, with 217 million daily active users on average[8]. It was intended as an urban lifestyle tool for friends, bringing together two subcultures: new media coding and radio scanning and dispatching. Originally the service was designed to work with the cellular administration of an SMS messaging service, thus the delivery constraints of text messages were the basis for the limit of the length a tweet could have (Rogers, 2013). Twitter later developed into a microblogging communication platform, allowing users to post "tweets" of up to 280 characters at a time, thus forcing users to drastically condense their thoughts and expression into a small, but saturated message. Twitter users can follow each other to see their tweets on their timelines, while tweets offer the possibility to mention other users, quote other tweets, or retweet a tweet to share it with their followers. Upon creating a Twitter account, users provide a screen name, a full name (real or fictitious), a location (although optional and not necessarily real) and a small autobiography (Freire & Graells-Garrido, 2019).

From a platform that was initially used for rather mundane and intimate messages, Twitter over time evolved into something bigger. As described by Rogers in 2013: "Twitter has

[7] https://twitter.com
[8] Twitter Q4 and Fiscal Year 2021 Letter to Shareholders, https://s22.q4cdn.com/826641620/files/doc_financials/2021/q4/Final-Q4'21-Shareholder-letter.pdf

evolved from a phatic and ambient intimacy machine, as Jack Dorsey envisaged it, to an event-following and news machine, as Biz Stone put it, when the Twitter tagline changed from “what are you doing?” to “what’s happening?” (Rogers, 2013). Twitter has proved itself as a useful tool during disasters and elections and became especially impactful during the San Diego fires, the Sichuan earthquake, the Mumbai terrorist attacks and the Hudson River plane landing (Rogers, 2013). When street demonstrations broke out after the presidential elections in Iran in June 2009 it was called “Iran’s Twitter Revolution” and in the wake of these events, Twitter was becoming a news source, sharing information from the ground, and replacing old media. Researchers started at that time to discover Twitter as a new object of study, de-banalizing it. The questions raised were about accuracy and professionalism in reporting, but also about the role the platform actually played in the presidential election protests that were taking place in Iran in 2009. Twitter then increasingly became an object of study as an emergency communication channel in cases of disasters and similar incidents as well as an event following and aid tool for the uprisings in the Middle East and beyond.

Twitter is especially attractive for research because of the relative ease with which tweets can be mined at a large scale (at least in real-time), as well as the inbuilt tools for analysis, including RT (retweets) for significant tweets, #hashtags for subject matter categorization, @replies as well as following/followers for network analysis (Rogers, 2013). The character limit, as well as the fact that every tweet has approximately the same length, also makes it suitable for textual analysis, such as co-wording analysis (Marres & Weltevrede, 2013).

Since tweets are very short by design (originally 140 characters long, expanded to 280 characters in 2017), the texts therein lack a broad conceptual framework that is usually present in longer forms of text. In addition, the language that tweets are written in is often quite colloquial, including many abbreviations, neologisms and slang, while the real-time, spontaneous nature of Twitter communication allows more room for spelling, grammatical and syntactic errors. All these particularities are expected to make the analysis of tweets much more complicated, as implementing formal dictionaries on the text can be more difficult and more attention needs to be paid on text-preprocessing. However, the brevity of tweets also means that Twitter users pay more attention into capturing the message more concisely and that emotion is more impactful, thus making the classification of tweets easier than their longer counterparts, blogs (Bermingham & Smeaton, 2010).

To investigate how Twitter users express themselves, it is important to understand the notion of the "Hashtag". Hashtags are characterized by the "#" character and typically highlight a trend among social media users. Hashtags are used for a multitude of reasons on Twitter and other social media platforms, including topic-marking, expressing emotion and attitudes, but hashtags should also be considered as a meaningful part of the message in the modern online discourse (Laucuka, 2018).

Using hashtags to explore a public opinion within Twitter can provide great insight as to what users who use a particular hashtag are concerned mostly about, as shown by a 2018 study on the hashtag #WorldEnvironmentDay that linked Twitter discussion around a particular hashtag with the main factors that concern the global population with respect to the sustainable development of the planet, public health, and the environment (Reyes-Menendez et al., 2018).

2.2 Twitter and research on refugees

There are a few key considerations while using Twitter data for human migration studies, especially for refugee movements, as Twitter penetration can vary greatly between countries, refugee populations do not always have network access, especially those en route and, perhaps most importantly, there are often security concerns within refugee communities, thus preventing them to broadly and openly use some social media.

A study from 2017 (Hübl et al., 2017), analyzing refugee migration patterns from the Middle East and Africa to Europe through geo-tagged tweets, suggests that only few refugees use Twitter, limiting the number of extracted travel trajectories to Europe, although the trends identified by the study also matched trends with data published by UNHCR, the UN Refugee Agency. Similarly, a study from 2020 (Aswad et al., 2020) focusing on how migrant communities figures, as detected from Twitter data, were conforming to official UN population data on migration found that social media can be misaligned with real-world data.

Researchers studying the refugee crisis in Europe, often use Twitter data for social media monitoring revolving around a particular event, be it via the use of hashtags or specific dates around that event. Often, these studies provide clearer outcomes due to their narrower data collection methodologies. A study from 2019 investigated hate speech on Twitter, when the Spanish government decided to allow the boat Aquarius, which carried asylum seekers, to disembark at the port of Valencia, after being rejected from other European ports. With over 24,000 tweets generated in just one week in June 2018, the study suggests that despite the overall majority of positive sentiments among the tweets about Aquarius, hate speech increased after the official announcement of the Government to accept the asylum seekers (Vázquez & Pérez, 2019).

Nerghes also uses hashtags to mine Twitter data on the days immediately following the death of Alan Kurdi[9] in September 2015, in order to investigate sentiments surrounding the words "refugee" and "migrant". The results of this paper clearly suggest a more positive tone on the tweets containing the word "refugee", while also containing the most neutrality compared to tweets containing the word "migrant", which typically bear more intense sentiment (Nerghes & Lee, 2018). Moreover, Nerghes uses network analysis to explore the relationship between the popularity of Twitter users and the polarity of their tweets, finding that both influential users and popular tweets portray less emotional intensity and even slightly less positivity (Nerghes & Lee, 2018).

Studies also clearly demonstrate how social media can complement traditional mainstream media and form a new media space, where narratives are created and transformed, especially during tragic events. A comparative study of mainstream media and Twitter in 2019 that focused again on historical tweets and publications on the period after the death of Alan Kurdi[9], used Latent Dirichlet Allocation (LDA) to compare the topics formed within the two media spaces (Nerghes & Lee, 2019). The researchers found that the two media environments can primarily act complementary to each other, rather than competing, although mainstream media portrayed a more neutral and broad discourse, compared to the more focused, highly sympathetic tone of social media, which also introduced new themes

[9] www.telegraph.co.uk/news/worldnews/europe/11841802/eu-migrant-crisis-refugee-boys-aylan-galip-kurdi.html

into the discussion (Nerghes & Lee, 2019). This latter finding is a rather expected result, as social media provide by nature an open platform for expression of their users and offer unlimited freedom to push public debates into new directions (Meraz & Papacharissi, 2013), contrary to the discussion we often witness in mainstream media.

Studies that focus on Twitter data mined to cover a broader time period are also able to identify specific events that took place during that period, mainly by identifying peaks in user activity and following the discussion during those times. A big data study from 2018 used Twitter data about refugees between October 2015 and May 2016 to investigate the discussion around refugees through hashtags and the networks of users talking about refugees, including relevant hubs (Siapera et al., 2018). Even though the scope of that study is broadly directed towards the refugee discussion on Twitter worldwide, researchers were able to identify major events that drove the discussion on Twitter, all of them originating in Europe. (Siapera et al., 2018) suggest that the publics that emerge from the refugee discussion revolve around established actors and narratives, while separation from the norm got lost among dominant tags and hubs. Perhaps that is due to the very broad scope of the data collection of that study, using the refugee keyword in seven languages and four additional hashtags, but their results are (not surprisingly) in-line with the nature of the Twitter platform itself, with tweets on trending topics often overpowering other narratives. (Siapera et al., 2018) also found that mainstream media were among the main social network hubs and that the refugee debate on Twitter revolves around two themes: security and racism on one hand and humanitarian response on the other.

Research presented in this chapter will act as the basis for the work in this dissertation.

3 Research methodology

This chapter will follow closely all the methodological steps of this dissertation, including the theoretical basis of the analysis.

3.1 Data mining

There are different approaches when it comes to mining Twitter data and choosing the one that is right for the issue at hand is the first big task that needs to be addressed for any research aiming to make use of Twitter data. For the purpose of this research, five separate mining methods were tried over time (May 2020 – December 2021) and are briefly overviewed here, along with their advantages and limitations. An important observation that needs to be noted is that some of the scraping methods described below might not be fully functional at the time this dissertation is published, as every time Twitter alters its front-end interface, the scrapers need to be updated as well and that is not always the case.

3.1.1 Twitter API

This method requires registering a developer account[10] with Twitter and going through a formal application process explaining the purpose and scope of the research, for which the data will be mined. The application process lasted 15 days from submission to approval, following which Twitter API credentials were granted (i.e., API key, API secret key, Bearer token, Access token and Access token secret). There are two types of Twitter APIs: Streaming API, offering tweet retrieval in real time, and Rest API, which is the chosen Twitter API for our purposes. The Tweepy[11] library was used to access Twitter's REST API with Python.

Despite the fact that Twitter's REST API works well and reliably once credentials are granted, there are some important limitations to its implementation:

---

[10] developer.twitter.com

[11] docs.tweepy.org/en/v4.10.0/api.html

- Requests are limited to 180 every 15 minutes and with a maximum number of 100 tweets per request[12].
- It is not possible to set a specific time-period for the requests, thus historical data cannot be accessed.
- Only Tweets written within the past seven days can be retrieved.

Given the fact that historical data were crucial for the purpose of this research, Twitter's REST API was eventually not used for the data mining of the tweet dataset used in this dissertation.

3.1.2 Twitterscraper[13]

Twitterscraper was a python library providing an alternative to Twitter's API for scraping Twitter data, which did not have any rate or date limitations and was able to retrieve a plethora of information from the social network. Twitterscraper was the first alternative data mining tool used for the purpose of this dissertation, but unfortunately it stopped working around September 2020 and has not been updated since. While at first, it was the preferred method of the author for data mining due to its reliability in mining consistent datasets, other methods had to be pursued eventually as results could not be duplicated after 2020.

3.1.3 Twint[14]

Twint is an advanced Twitter scraping & OSINT tool, written in Python, that does not use Twitter's API, allowing users to scrape without rate limitations. At the same time, Twint can overcome the seven-day limitation that Twitter's API has for historical data by scraping from the most recent tweets going backwards to older ones.

During the data mining efforts of this research, the most common issues encountered while using Twint was the frequent IP and device bans, which made it particularly difficult to scrape consistently, coupled with the malfunction of date parameters at the time of scraping.

[12] developer.twitter.com/en/docs/twitter-api/v1/rate-limits
[13] github.com/taspinar/twitterscraper
[14] github.com/twintproject/twint

For these reasons, Twint was eventually abandoned as a data mining strategy, as the needed tweets were dating as far back as 2015 and this method proved to be too time consuming.

3.1.4 Selenium[15]

Selenium is an open-source web browser automation tool that allows users to simulate web navigation operations that would otherwise be too time-consuming for a human to perform. One of the uses for Selenium's WebDriver is web scraping, including for Twitter data, but that method does require some understanding of basic HTML language, along with real Twitter credentials.

One of the challenges faced while using Selenium to scrape Twitter data for this dissertation was that the webpage needed to fully load before Selenium attempted to locate the element it was programmed to search for, otherwise an error message would be prompted, and the process would need to initiate again from the beginning. Due to the fact that we needed to go back to 2015 to search for relevant tweets for this research and, thus, Selenium needed to be operational for a significant number of hours, when the internet connection was not consistent, Selenium would crash. At the same time, Twitter does temporarily block accounts with suspicious behavior, so the web scraping needed to be postponed a few times. Selenium was eventually abandoned as the chosen tool for the purposes of this dissertation, but it proved to be very useful at a time when many of the more well-known scrapers did not function (spring/summer 2021).

3.1.5 Snscrape [16]

Snscrape is a scraper for social networking services, including Twitter, and can retrieve tweets that include a large list of attributes, based on a given keyword, hashtag or user profile. Snscrape can overcome the seven-day limitation that Twitter's API has for historical data and can theoretically retrieve up to 100k tweets per day.

[15] https://www.selenium.dev/
[16] github.com/JustAnotherArchivist/snscrape

Taking into consideration all five Twitter scraping methods tried for the purposes of this research, the only viable choice at the time of the final data mining (January 2022) that was also time efficient was Snscrape, as it overcame many of the restrictions that Twitter's API imposes, while also retrieving consistent data each time it was used.

Snscrape retrieves a variety of information for a single tweet, not just its text. A brief overview of the most important attributes available through Snscrape is shown in the table below.

Table 1: Snscrape attributes

Attribute	Description
url	A permanent link directing to the location of the tweet
date	The date the tweet was posted
renderedContent	The full text of the tweet
id	The unique identification number of the tweet
user	This is an object that contains further information on the user posting the tweet, such as: username, displayname, id, description, verified, created, followersCount, friendsCount, location, linkUrl, profileImageUrl and a few more
replyCount	How many replies the tweet received
retweetCount	How many times the tweet was retweeted
likeCount	How many likes the tweet received
quoteCount	How many times the tweet was quoted and replied
lang	The language of the tweet, as generated by Twitter
source	Information on the device that was used for the tweet to be posted
retweetedTweet	The content of the retweeted tweet (if any)
quotedTweet	The content of the quoted tweet (if any)
mentionedUsers	User objects of any mentioned user within the tweet

3.2 Keywords and time frame

As the scope of this dissertation is to investigate and understand some of the components shaping the public opinion of Twitter users about the refugee crisis in Greece, it is important that the time aspect of the data mining is as broad as possible. In addition, the hashtag that should be used to filter which tweets are to be extracted needs to be an inclusive and descriptive one that is used not only by local people and media focused on a specific demographic, but also international media, organizations involved with refugees and politicians alike. Taking those constraints into consideration, the hashtag chosen is #RefugeesGR and the time frame spans from April 2015, when the account "NoBorders" first joined Twitter and the hashtag #RefugeesGR was first used, until the end of December 2021.

Due to the fact that a specific hashtag was used to retrieve the dataset, this body of tweet data is not exhaustive when it comes to tweets discussing about the refugee crisis in Greece over the time period 2015 – 2021, albeit it does offer a deep insight due to its extended use by regular Twitter users, not-for-profit and non-governmental organizations, public figures, politicians and international organizations alike.

3.3 Data cleaning and preprocessing

Data cleaning is an important part of natural language processing, especially when it comes to texts scraped from microblogging social networks, as tweets tend to be unstructured, contain a plethora of special characters as well as links, hashtags and mentions. The process of data cleaning is a lengthy task that needs to be performed very early in the analysis stage and also be reviewed and refined multiple times, so that data quality issues can be efficiently resolved prior to any analysis. The key steps for data cleaning and preprocessing in natural language processing are to remove punctuation, remove stopwords, normalizing the text, stemming or lemmatizing the words and tokenizing the text.

Going forward with the tweet dataset at hand, the focus of the linguistic part of this work will be concentrated to a specific language among those the Twitter users were communicating in while using the hashtag #RefugeesGR and that language was chosen to be "English", due to the volume of the tweets and the widespread sectoral variety of Twitter users using it, both in Greece and internationally.

3.3.1 Removing "noise" and stopwords

For the purposes of this dissertation, the Python library 're' was used to write and use a function containing regular expression (*regex*) operations. As a first step in cleaning the English tweet dataset, all punctuation, hyperlinks, numbers, hashtags, mentions, were identified and removed and the text was all converted to lower case characters.

Stopwords are a set of commonly used words that typically include articles, prepositions, conjunctions and pronouns. The point of removing stopwords in the data preprocessing step is so the model can focus on words that have a more significant meaning, since stopwords rarely alter the meaning of a sentence and have little to contribute to the context. Blei et al. also suggest that common stopwords should be removed prior to running Latent Dirichlet Allocation (Blei et al., 2003). In addition to removing stopwords for the purposes of topic modelling in this dissertation, a set of very common words, collection words and other interference words were removed in order to create wordclouds and most-common-word charts. For the purposes of this dissertation, stopwords were filtered using the nltk[17] library.

3.3.2 Text normalization

Text normalization is a tool used to transform unstructured text, in order to bring it closer to a standard form, and it usually includes *stemming* or *lemmatization* and *tokenization*.

Both stemming and lemmatization techniques are ultimately used to reduce a word to a more common base form, but there are key differences between the two. *Stemming* is a more basic and straightforward process of normalizing a word, as it essentially removes letters from the end of the word until the basic stem remains. This method does work well in the majority of cases, especially in the English language, but it does not account for grammatical exceptions.

[17] www.nltk.org/

Lemmatization on the other hand considers the full vocabulary of a language in order to reduce each word to its proper base form.

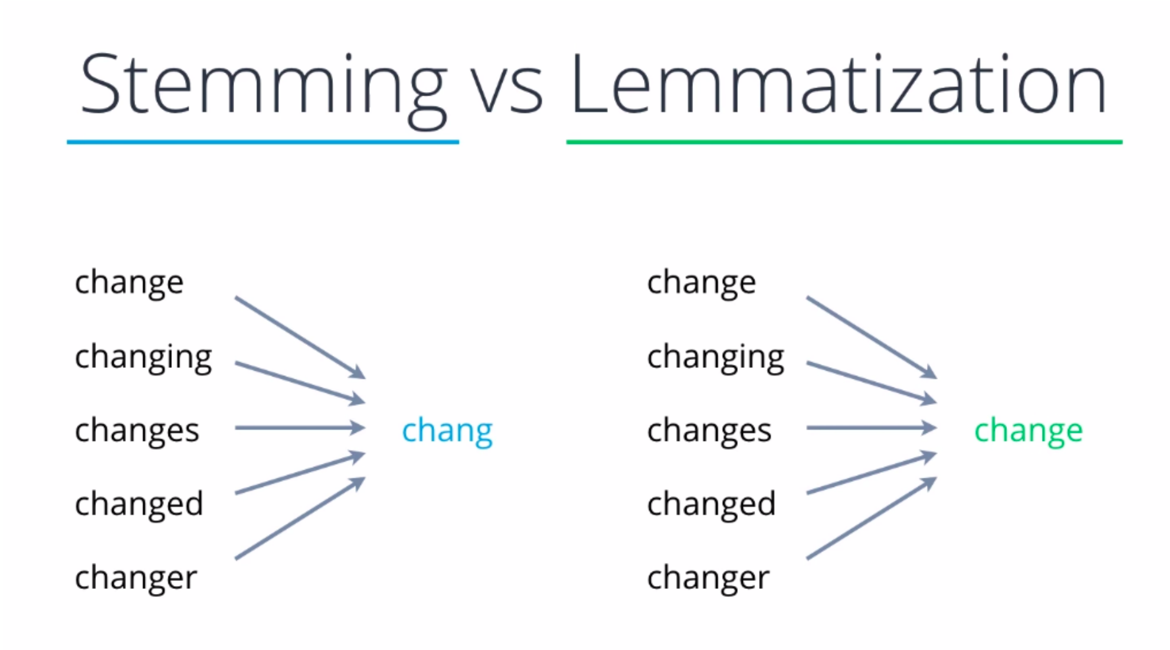

Figure 3-1: Example on stemming and lemmatization application in the same words[18]

For the reasons described above, lemmatization will be used in this dissertation for standardizing words into their proper form, as the tweet data will be more consistent in that manner and the quality of the results is expected to be better. The chosen method to achieve lemmatization in this work is the library spacy[19].

Tokenization is the process of partitioning a raw text into smaller parts, for example breaking a paragraph into sentences, a sentence into words and a word into individual characters. The tweets in this dataset were separated into individual *tokens*, i.e., words, for the purposes of further analysis.

[18] Image by Prateek Sawhney (medium.com/geekculture/introduction-to-stemming-and-lemmatization-nlp-3b7617d84e65)
[19] spacy.io/api/lemmatizer

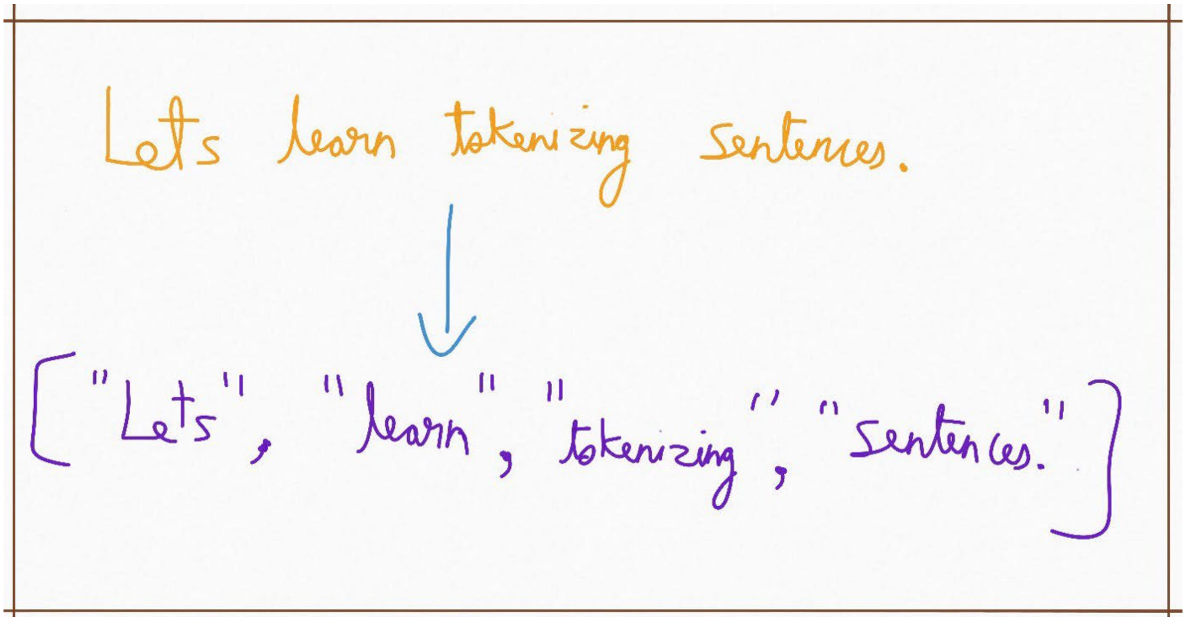

Figure 3-2: Example [20]of tokenization method for sentences

3.4 Yearly datasets and term frequency

The purpose of scraping tweets on the refugee situation in Greece for such a long period of time, namely 2015 – 2021, is mainly to be able to understand some key changes that occurred over time in the way Twitter users express their views, opinions and concerns on the matter. Thus, the tweet dataset was portioned into yearly datasets, each corresponding to a full year (1 January to 31 December), with the exception of 2015, because tweets with the hashtag #RefugeesGR first appeared in April of that year.

In order to compare the immediate similarities and differences between these yearly tweet datasets, term frequency charts were created, both for individual hashtags and words, along with wordclouds, as the frequency of terms is more presentable in this manner.

[20] Image by Dhaval Thakur (python.plainenglish.io/how-to-tokenize-sentences-without-using-any-nlp-library-in-python-a381b75f7d22)

3.5 Topic modelling

Topic modelling is an important tool for natural language processing in machine learning, that uses algorithms in order to identify the main themes in a collection of documents. The assumption pathway which topic modelling follows is that documents are a collection of topics and topics are a collection of tokens (e.g., words), so that, by identifying a series of tokens relevant to each other, an overarching theme can emerge. A topic modelling algorithm, which most commonly falls under the category of unsupervised learning, can produce a set of topics that capture the underlying concepts of the whole corpus, by outlining a probability distribution over a set of co-occurring tokens. Topic modelling can be easily comprehended when compared to clustering methodologies, i.e., clusters of tokens that formulate the topics.

In order to be able to apply a topic modelling algorithm to the data at hand, a *corpus* needs to be established first. Following that step, a *dictionary* should be created, giving each word in the corpus a unique identifier and then a *document-term matrix*.

3.5.1 Document term matrix and bag-of-words

Following the data cleaning and pre-processing step of the process, the outcome is a collection of tweets, all separated into tokenized and lemmatized terms, constituting the *corpus*. The next step is to understand how frequently each term is present within each tweet, constructing a *document-term matrix*.

To achieve that, a *bag-of-words* model is used, specifically Gensim's doc2bow for the purposes of this dissertation, as a method of capturing the terms in the documents of the corpus. The bag-of-words is a vector space model, where each document is expressed through a vector in an n-dimensional space, where n is the number of unique terms across the corpus vocabulary. The bag-of-words assumption is based on the principle that the order of the words in a given document is not important and can be overlooked, as is the order of the documents themselves. Ultimately, the bag-of-words model is a representation of the documents and their texts, where only the counts of each word are important.

An early notion of the bag-of-words concept, although made in linguistic terms, was presented by Harris in 1954 in his work on the distributional structure of language: “And this stock of combinations of elements becomes a factor in the way later choices are made […] for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use” (Harris, 1954).

3.5.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic topic modelling technique proposed by David Blei, Andrew Ng, and Michael Jordan in 2003 (Blei et al., 2003). The concept behind the LDA model is that documents can be described using a distribution of topics. The topics, using the bag-of-words assumption described above, are formally defined as a distribution over a fixed vocabulary (Lafferty & Blei, 2009).

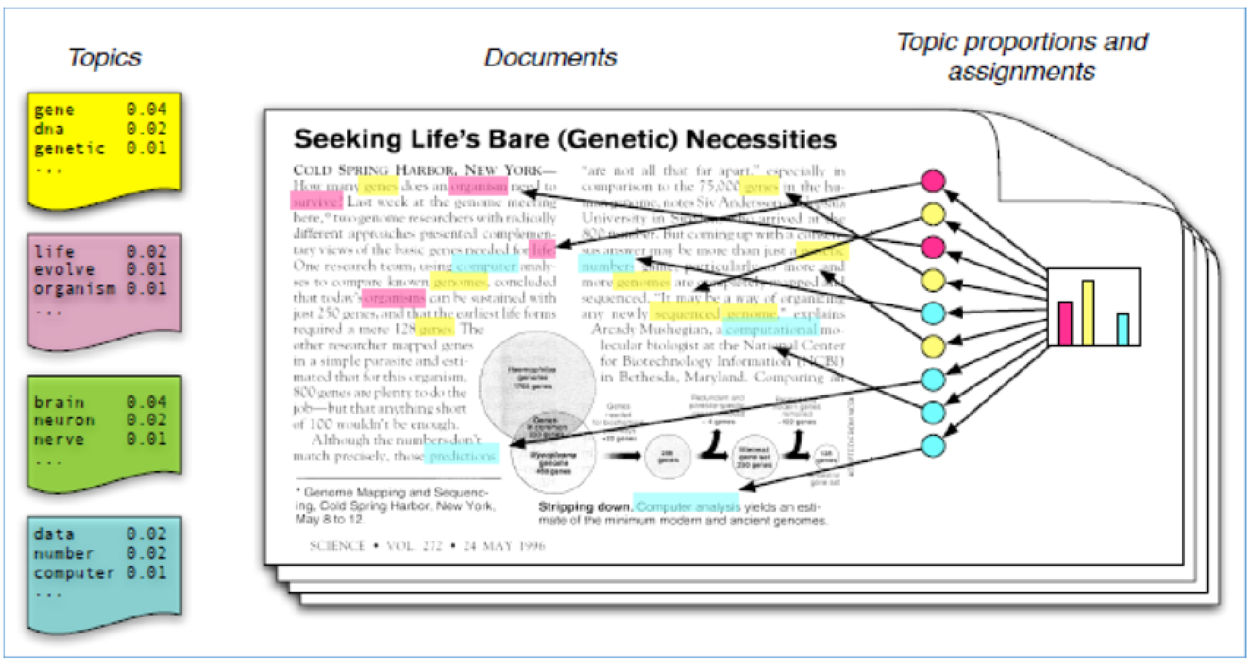

Figure 3-3: The intuitions behind latent Dirichlet allocation.(Blei, 2012)

In Figure 3-3 (Blei, 2012), Blei showcases an example of probabilistic topic models in a visual manner. Taking an article on "Seeking Life's Bare (Genetics) Necessities", Blei creates an illustration with each highlighted word corresponding to the same-colored coin on the right. The coin distribution is also presented in a histogram of the topics in the article (also on the right), while on the left side the topics are displayed as building blocks, with each word having a probability assigned to it.

To better understand how LDA works, it is helpful to study the plate diagram of LDA in Figure 3-4 below (Blei et al., 2003), where:

- the corpus consists of M documents, which in turn consist of N words
- w – a specific word among N
- α – the topic distributions for each document in the corpus
- θ – the topic distribution specifically for a document among M
- z – a specific topic
- η – the word distribution for each topic
- β – the word distribution specifically for a topic
- k – the number of topics

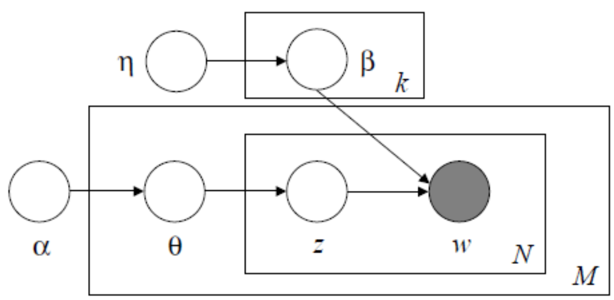

Figure 3-4: Graphical model representation of smoothed LDA (Blei et al., 2003)

Following along the LDA plate model in Figure 3-4 above, there is an α distribution of topics for all documents which defines the θ distribution of a specific document among M. Going into the inner plate, there are N words in that document M and each word is assigned to a topic z. Looking into the plate on the top of the figure, there is an η word distribution for each topic, which β is a part of. Using the Dirichlet distribution (multivariate generalization of the Beta distribution) β assigns k words for each topic based on which topics are already in the particular document and how many times this particular word was assigned a specific topic.

By repeating the process for all words within *N* and for all documents within *M*, there is now a topic distribution for the whole corpus.

LDA falls under the unsupervised machine learning algorithms that learn patterns from unlabeled data, however, it does require the preselection of the parameter *k*, referring to the total number of topics the corpus should consist of. This parameter *k* can vary greatly depending on the corpus at hand, but there are some qualitative evaluation metrics that can help narrow down this number, specifically *perplexity* and *coherence*.

The most common way to evaluate a probabilistic model is to measure the log-likelihood of a held-out test set. However, (Chang et al., 2009) show that predictive likelihood (or similarly, *perplexity*) and human judgment are often not correlated, and even sometimes slightly anti-correlated. Therefore, optimizing the number of topics of the LDA model based on the perplexity measure is not the approach followed in this dissertation.

The *coherence* measure on the other hand, provides a score for a single topic by measuring the degree of semantic similarity between high scoring words within the same topic. These measurements help distinguish between topics that are semantically interpretable and topics that are merely artifacts of statistical inference (Stevens et al., 2012). There are a few different options to calculate coherence measures (such as *c_v*, *c_p*, *c_uci*, *c_umass*, *c_npmi*, *c_a*) and in this body of work the measure *c_v* is used. The measure *c_v* is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity (Mifrah, 2020).

In addition to any qualitative evaluation metrics that can help with the selection of the optimal number of *k* topics in the LDA model, it is also still quite important to include human judgement into these decisions as natural language is messy, ambiguous and full of subjective interpretation, much more when it comes to microblogging texts such as those scraped from Twitter. So, in order to fully evaluate the topic modelling algorithm, there is still need for understanding of the corpus as a whole, obtaining the most common words in the corpus and interpreting the topics produced by the model against real-world information and judgment.

For the purposes of topic modelling in this dissertation, the python library Gensim[21] was used. It is also worth mentioning that an attempt to use the nltk library for topic modelling was made (working quite well no less), however, Gensim was eventually the chosen one due to its compatibility with the preferred visualization method (described below).

3.5.3 Topic modelling visualization with LDAvis

Once the topic modelling algorithm is selected, tuned and implemented, the next step is interpretation of its results. This task is facilitated with the use of a visualization tool, which, for the purposes of this dissertation, will be LDAvis.

LDAvis (Sievert & Shirley, 2014) is a web-based interactive visualization of topics, that has become one of the most popular tools for the presentation of topics when using an LDA model, as it allows for both a global view of the topics in the model, including how far apart they are from each other topologically, and a deep inspection of the terms most highly associated with each individual topic.

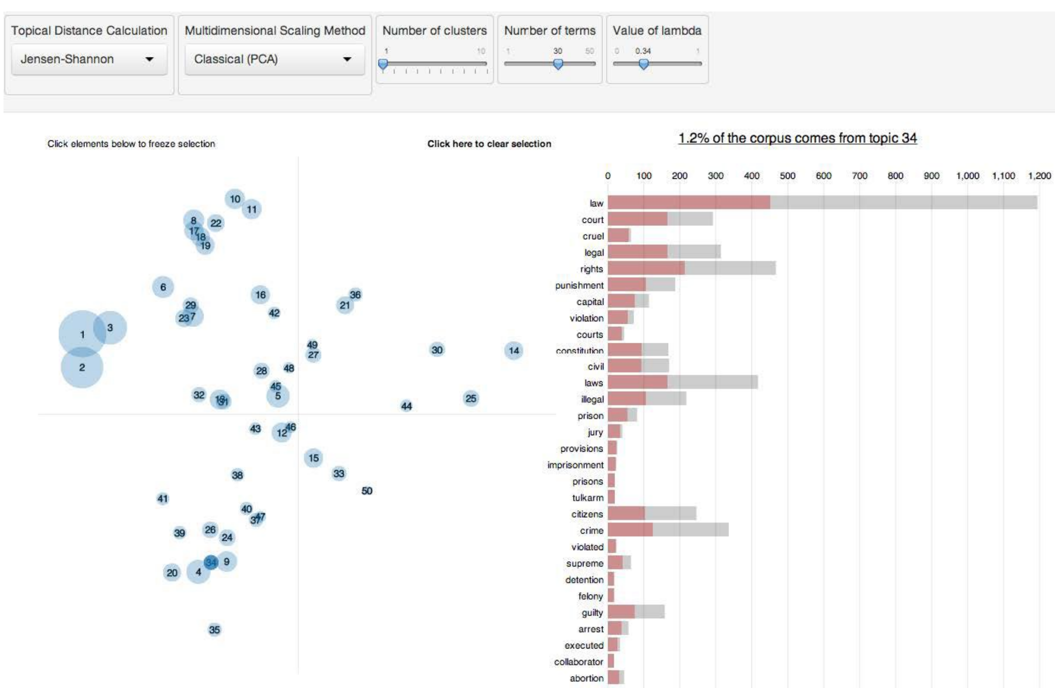

Figure 3-5: The layout of LDAvis (Sievert & Shirley, 2014)

[21] pypi.org/project/gensim

In Figure 3-5 (Sievert & Shirley, 2014), a sample output of LDAvis can be observed. Each of the bubbles on the left side of the visualization represents a topic and the size of the bubbles represent the percentage of documents in the corpus that are associated with that topic. The position of the bubbles reflects how closely correlated the topics are with each other. The bars on the right side of the visualization represent the frequency of the terms, which change with the selection of bubbles on the left.
To achieve a similar result on the tweet dataset scraped for this dissertation, the pyLDAvis library[22] will be utilized and the results are showcased in Chapter 4.4.

3.6 Social network analysis

Social network analysis (SNA) originates from graph theory, a branch of mathematics used to model -pairwise- relations between objects. A social network structure is formed when connections are created among social actors such as individuals and organizations (Wasserman & Faust, 1994). Studying these networks and the connections between the social actors can help with understanding the dynamics between the participants of the networks and also how communities are formed and interact therein. As described by (Wetherell et al., 1994) in a broad manner, social network analysis (1) conceptualizes social structure as a network with ties connecting members and channeling resources, (2) focuses on the characteristics of ties rather than on the characteristics of the individual members, and (3) views communities as 'personal communities', that is, as networks of individual relations that people foster, maintain, and use in the course of their daily lives.

Now, to understand what a social network structure might look like, the first step would be to conceptualize the two main entities to be found within a graph, namely:

1. *Nodes* (or vertices)
2. and *Edges* (or lines)

The edges in a graph connect two nodes that have a certain relationship and can be directed, given that the relationship between two nodes follows a specific orientation, or undirected,

[22] pyldavis.readthedocs.io/en/latest/index.html

when the edges do not follow any orientation. Two nodes might not be directly connected to each other in a graph, however, there often is a certain path that can be followed to reach one node from another (with the exception of a disconnected graph or isolated nodes).

For the purposes of this dissertation, the *nodes* in our network represent Twitter users and the *edges* represent interactions between them. This concept can be easily understood through the visualization in Figure 3-6 below.

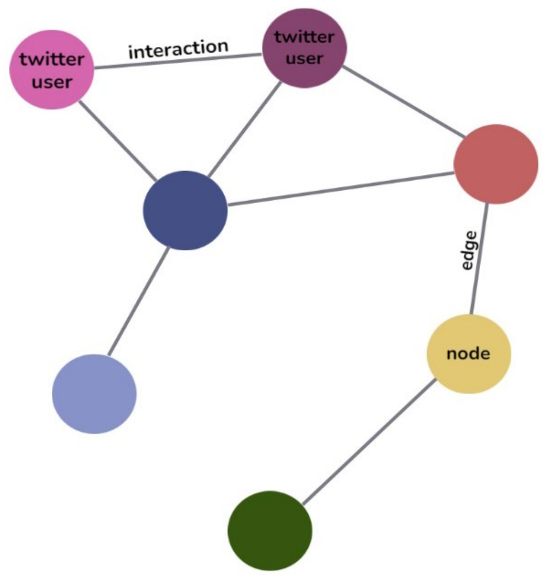

Figure 3-6: Author's example of a graph within the Twitter user framework

Interactions withing the Twitter framework can be retweets, quotes, replies, mentions or followers and these are generally actions that can be represented by a directed network, where for example one Twitter user is mentioning another user in their tweet.

The consensus here is that there is an implied relationship between two Twitter users interacting, on the basis of one retweeting, quoting, replying, following or mentioning the other user. Through this implied relationship, it is possible to identify the most influential twitter users, the key roles some of them play within the network and the communities formulated by groups of users, as per their interactions with each other.

3.6.1 Network measures

To better understand what these relationships might look like, there are some key notions and measures in a network that need to be identified first.

- The *density* of the network is a measure of how closely connected its nodes are. It is calculated with a simple ratio of the present edges in a network to all the possible edges in the network and it ranges from 0 to 1.
- *Reciprocity* is a measure used for directed networks and represents the likelihood of nodes to be mutually linked to one another. For the purposes of a Twitter network, reciprocity can be described as two users following each other.
- *Degree* is a measure of centrality, and it refers to the number of connections a node has within the network, a simple sum of its edges. For directed networks, this measure includes both incoming and outgoing edges.
- *Eigenvector centrality* can be seen as an expansion of the degree concept above, as it incorporates not only the edges that a node has, but also the edges that the node's neighbors have to its measure, and it ranges from 0 to 1. This approach factors the entire network into the ranking of the nodes, and it is a helpful measure to understand how influential a node really is by examining its role within its community of neighbors.
- *Betweenness centrality* differs from degree and eigenvector centrality measures, as it is only based upon the number of shortest paths that the node is part of within the network. The measure also ranges from 0 to 1 and it provides an insight on whether a particular node can be crucial in connecting two clusters of nodes.

Based on the measures identified above, it's important to also mention two roles that can be assigned to particular nodes, that will prove useful in analyzing the Twitter user network later.

- A *Hub* represents a central user within a network, someone important among the community, who has a high degree centrality and/or high eigenvalue centrality.
- A *Broker* also represents an essential user for the flow of information within a network, as their node connects two parts of the network that would otherwise be disconnected, which means that this node has a high betweenness centrality.

An important question when it comes to analyzing a social network, especially one consisting of Twitter users, is whether it presents a uniform body of similar people or is essentially a collection of smaller *communities* that have more in common within their particular groups. Essentially, the identification of communities within a social network is based on the notion of homophily. Homophily theory, usually referenced together with the proverb "birds of a feather flock together", was first introduced in 1954 by sociologists Paul Lazarsfeld and Robert Merton in their study of friendship processes. In a social network, homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people and implies that distance in terms of social characteristics translates into network distance (McPherson et al., 2001). Thus, as homophily suggests, individuals often form social networks with people who are similar to them.

A very popular method to calculate communities within a network is *modularity* and it is based on how dense the different clusters are when compared to nodes outside their cluster. It is often easier to partition more fragmented networks than dense ones, as the number of actual connections compared to the number of all possible connections is relatively low. For the purposes of this dissertation, the python package networkx[23] was used both to generate the network and calculate its measures, while for community detection and partitioning, an extra python package was used, complementary to networkx, namely the community API[24].

3.6.2 Network visualization

Following the generation of the network of Twitter users, the visualization aspect needs to be addressed. Graph visualization is an important step in the network analysis process, as it allows for quick and intuitive identification of patterns, trends and outliers[25]. While there are a number of good options available for network visualization, such as NodeXL, the Gephi software (Bastian et al., 2009) was chosen by the author, as it can accommodate large graphs easily, offers a variety of algorithms to adjust the graph's layout and it also works well with networkx, as the latter can output graphs in Gephi's .gexf format with pre-calculated

[23] networkx.org
[24] perso.crans.org/aynaud/communities/index.html
[25] Cambridge Intelligence white paper on network visualization: cambridge-intelligence.com/wp-content/uploads/2021/01/Graph-visualization-WP-compressed.pdf

measures for centrality and modularity (and also many more as needed). Gephi[26] is ultimately an open-source software for visualizing and analyzing large networks, using a 3D render engine.

Visualizing a network is never trivial and comes with a multitude of choices regarding the sizing of the nodes, the coloring of the nodes and edges and the preferred layout. For the purposes of this dissertation, the eigenvector centrality measure was chosen to size the nodes, as it takes into account the network as a whole and it can give a better perspective on importance, especially when users vary significantly in popularity within a network. In addition, the nodes were colored based on the modularity measure of the community finding algorithm mentioned above and the colors were solely assigned to communities with 50 Twitter users or above, in order to keep the visualization readable.

Modelling the layout[27] of the network was a very time-consuming task, as some algorithms take significantly longer to compute, and others crashed altogether. After hundreds of attempts, two modelling algorithms were chosen in combination, in order for the network visualization to have communities somewhat separated from each other, hubs clearly visible and nodes not overlapping with each other. The result was a product of multiple trial-and-error attempts at finding the suitable combination of layouts that complemented each other's results. At first, the OpenOrd layout (Martin et al., 2011) was used, as it is really efficient in distinguishing clusters from one another and it works well for larger graphs. However, using OpenOrd on Gephi proved to produce a visualization with a very high number of overlapping nodes, which made it quite difficult to distinguish the individual labels of each node. So, as a second step, the more popular ForceAtlas2 layout (Jacomy et al., 2014) was used, which offers a variety of settings that allow for a precise arrangement of the visualization of the network. This double layout methodology allowed the author to solve the issue of overlapping nodes while also having hubs more clearly visible.

This chapter covered the theoretical basis and the methodological approach followed in the analysis of the dataset minded from Twitter, while the findings of the analysis will be presented next in Chapter 4.

---

[26] gephi.org

[27] gephi.org/tutorials/gephi-tutorial-layouts.pdf

4 Data analysis and findings

This chapter of the dissertation will cover the output of the data analysis, starting with the presentation of some summary statistical findings derived from the mined dataset of tweets that include the hashtag #RefugeesGR, along with the accounts posting them.

4.1 Data overview

The final dataset mined from Twitter contains 116,530 tweets, dating from April 2015 to December 2021 and includes tweets in 44 languages. The main languages of the tweets using the hashtag #RefugeesGR are:

1. Greek, with a total of 48,488 tweets,
2. English, with a total of 41,059 tweets,
3. German, with a total of 6,758 tweets,
4. and Spanish, with a total of 3,721 tweets.

French, Italian, Turkish, Dutch, Farsi/Dari and Portuguese were also present in the dataset, albeit with less than 1,000 tweets per language. In the final dataset, there were also 13,613 tweets, the language of which was marked as 'undefined' by Twitter and included tweets either containing text in multiple languages or solely hashtags.

The volume of the extracted tweets and how it fluctuates over the time during the 2015 – 2021 period is summarized both in Table 2 (numerically) and Figure 4-1 (visually) below, where the volume of the whole dataset is shown, as well as the volume of tweets in Greek and English separately.

Table 2: Number of tweets using the hashtag #RefugeesGR per year and dominant language.

Out[16]:

	year	tweets	GRtweets	ENGtweets
0	2015	25011	9475	10141
1	2016	43871	17084	16680
2	2017	12441	3441	6701
3	2018	6047	1920	3074
4	2019	6576	2367	2538
5	2020	16325	4960	6505
6	2021	6259	1812	2849

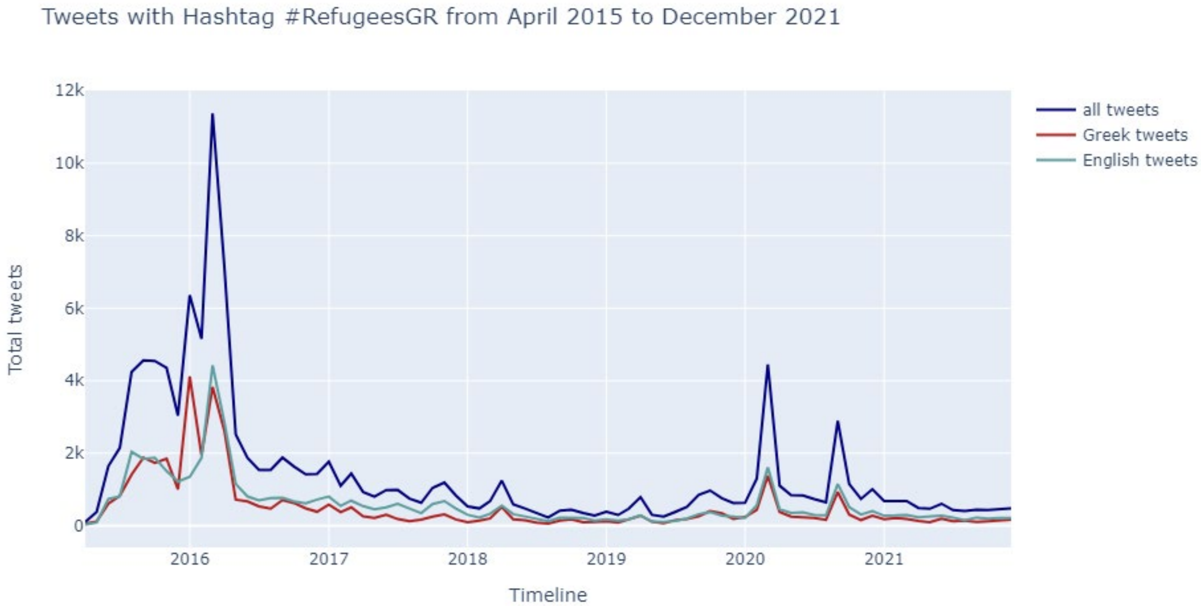

Figure 4-1: Tweet volume timeline (May 2015 - December 2021) per language

The hashtag #RefugeesGR began being widely used amongst Twitter users in June 2015 and became very popular in August 2015, with a monthly average exceeding 4,000 tweets per month for the remainder of 2015. As Figure 4-1 suggests, the monthly average of tweets using the hashtag #RefugeesGR continues to grow until March 2016, when there is a sudden growth of more than double the monthly average up to that point. Following that peak in 2016, the hashtag usage slowly levels down during the last six months of 2016, with 2017 having a monthly average of some 1,000 tweets. Both 2018 and 2019 portray similar trends in the usage of the hashtag #RefugeesGR on Twitter, with a monthly average of some 500 tweets and no more than three peaks per year. However, the hashtag once again gains wide popularity in 2020 with two significant peaks of about 4,500 and 3,000 tweets respectively, before falling back to a monthly average of 500 tweets during 2021. The volume of tweets per year and the events that sparked the peaks described above will be discussed further in Chapter 4.2.

4.1.1 Twitter users

Moving on from the volume of tweets to who is posting them, there are a total of 11,754 distinct users within the mined dataset, who have posted tweets using the hashtag #RefugeesGR during the years 2015 – 2021.

A first step towards understanding the characteristics of those 11,754 Twitter users is to investigate the attributes that separate them, namely: the number of followers, the number of friends[28], the activity volume and the verification status[29]. Figure 4-2 depicts a visualization of the distribution of Twitter users posting about #RefugeesGR, based on the number of followers, number of friends, number of posts and verification status of their account. Looking at Figure 4-2, there is a clear correlation between the number of followers a user has and their verification status, albeit with one distinct outlier having more than five million followers. In addition, Twitter users who follow more than 10,000 Twitter accounts are mainly non-verified. The amount of activity per user account, however, does not seem to provide any meaningful conclusion here, other than the fact that verified accounts do not portray extreme posting behavior overall.

[28] The number of "friends" here refers to the Twitter accounts that a specific user follows: developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/overview#:~:text=Friends%20%2D%20we%20refer%20to%20%22friends,that%20follow%20a%20specific%20user

[29] The verification badge for a Twitter user is explained further within Chapter 4.1.1

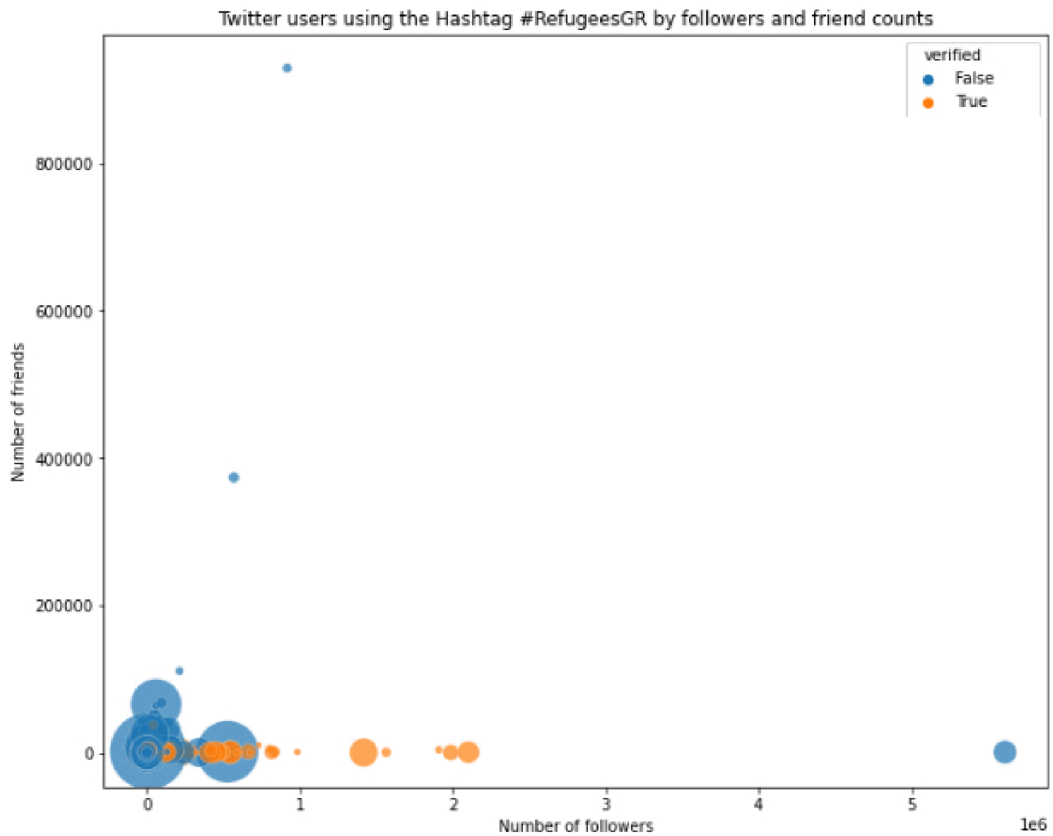

Figure 4-2: Distribution of Twitter users based on followers, friends, posts and verification status

Focusing now on Twitter accounts which post about #RefugeesGR and have the most following, Figure 4-3 visualizes the list of top twenty accounts based on the number of followers and Figure 4-4 provides the profile snapshots of the top nine most followed accounts. Figure 4-4 is of particular interest, as it showcases that seven out the nine most followed accounts are -in fact- verified accounts and, among them, three belong to news organizations, three to international humanitarian organizations and one to the European Union.

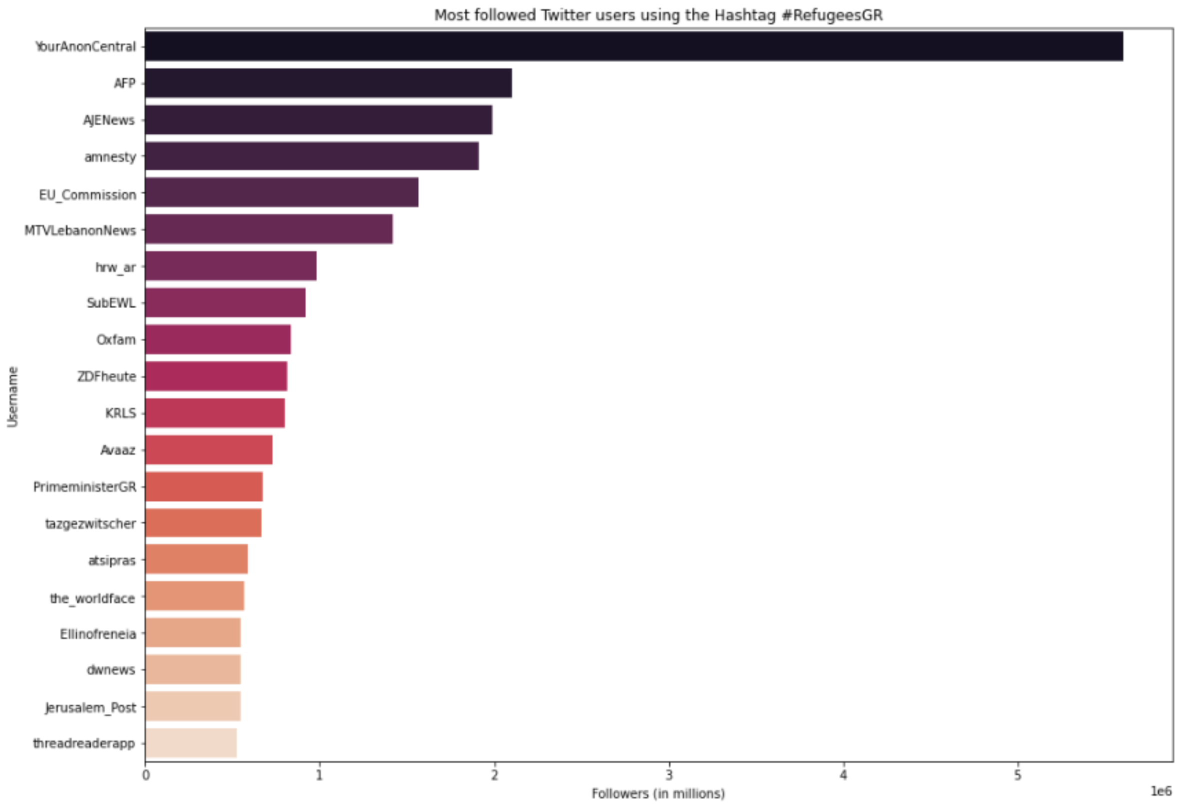

Figure 4-3: Top twenty Twitter accounts posting about #RefugeesGR with the most followers

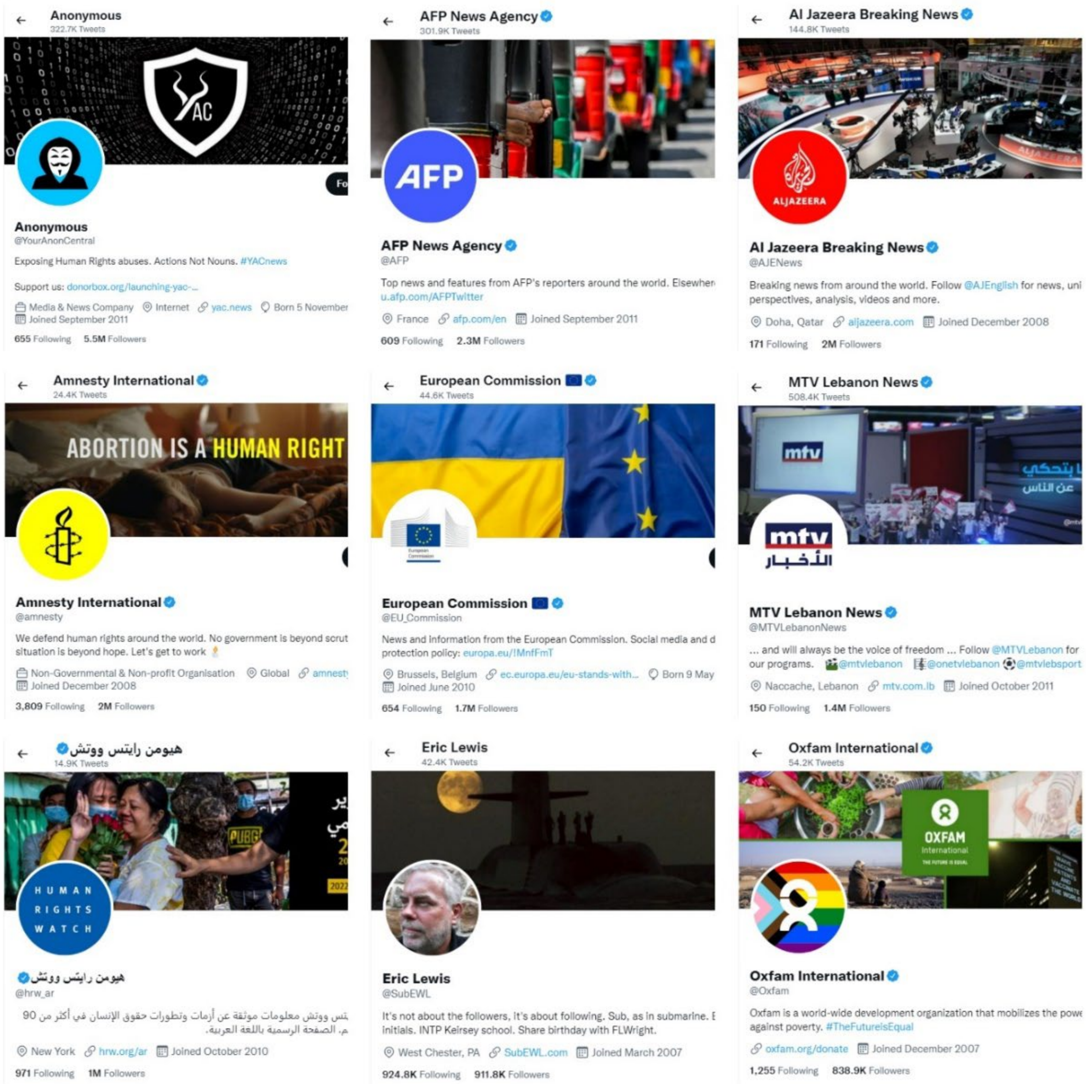

Figure 4-4: Twitter users using the hashtag #RefugeesGR with the most followers

Continuing with the Twitter accounts which posted the most about #RefugeesGR over the period 2015 - 2021, Figure 4-5 depicts the list of top twenty accounts based on number of posts that include the hashtag #RefugeesGR and Figure 4-6 provides the profile snapshots of the top nine[30] of those accounts. Looking closely at Figure 4-6, there are three accounts relating to news organizations and private journalists (two of which are also verified), two

[30] The account that was originally in the eighth place (@Chara_fc) has been disabled and it was replaced in the collage of Figure 4-6 with the account in the ninth place.

accounts affiliated with human rights and civil society organizations and three activist accounts.

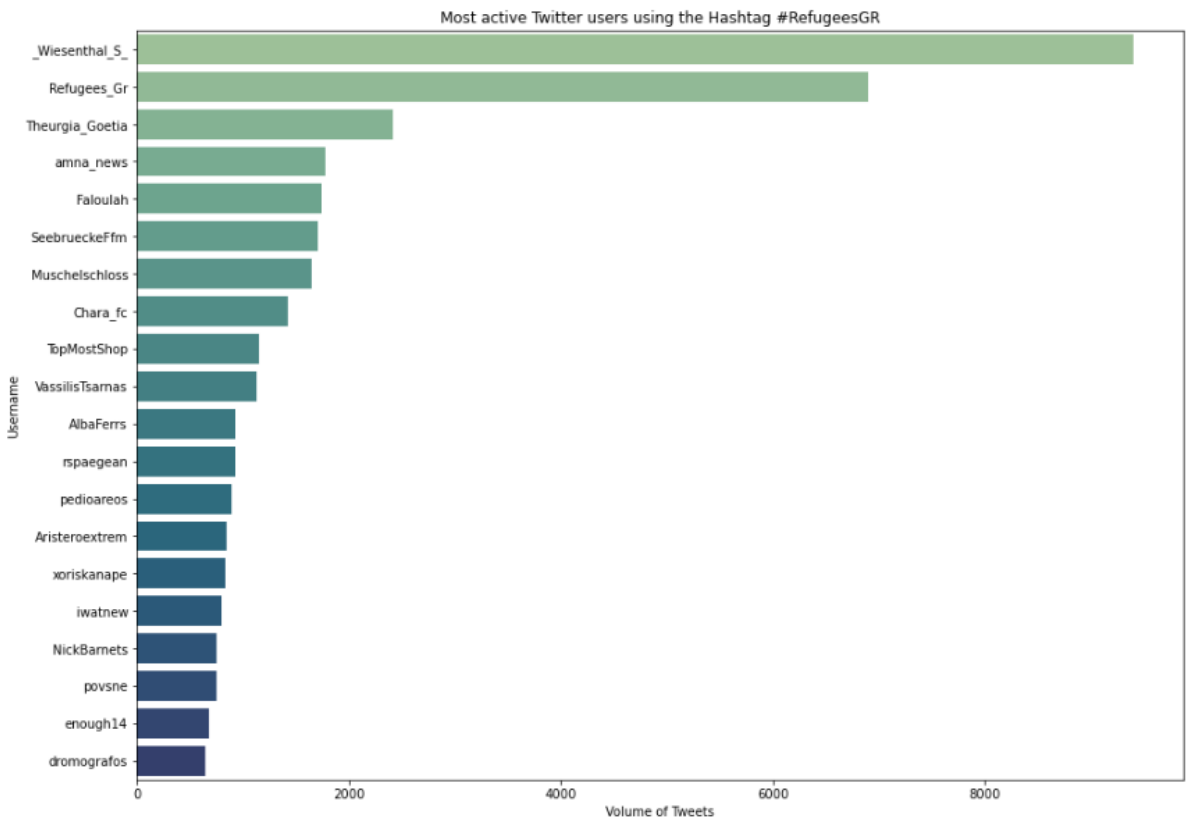

Figure 4-5: Top twenty most active Twitter accounts posting about #RefugeesGR

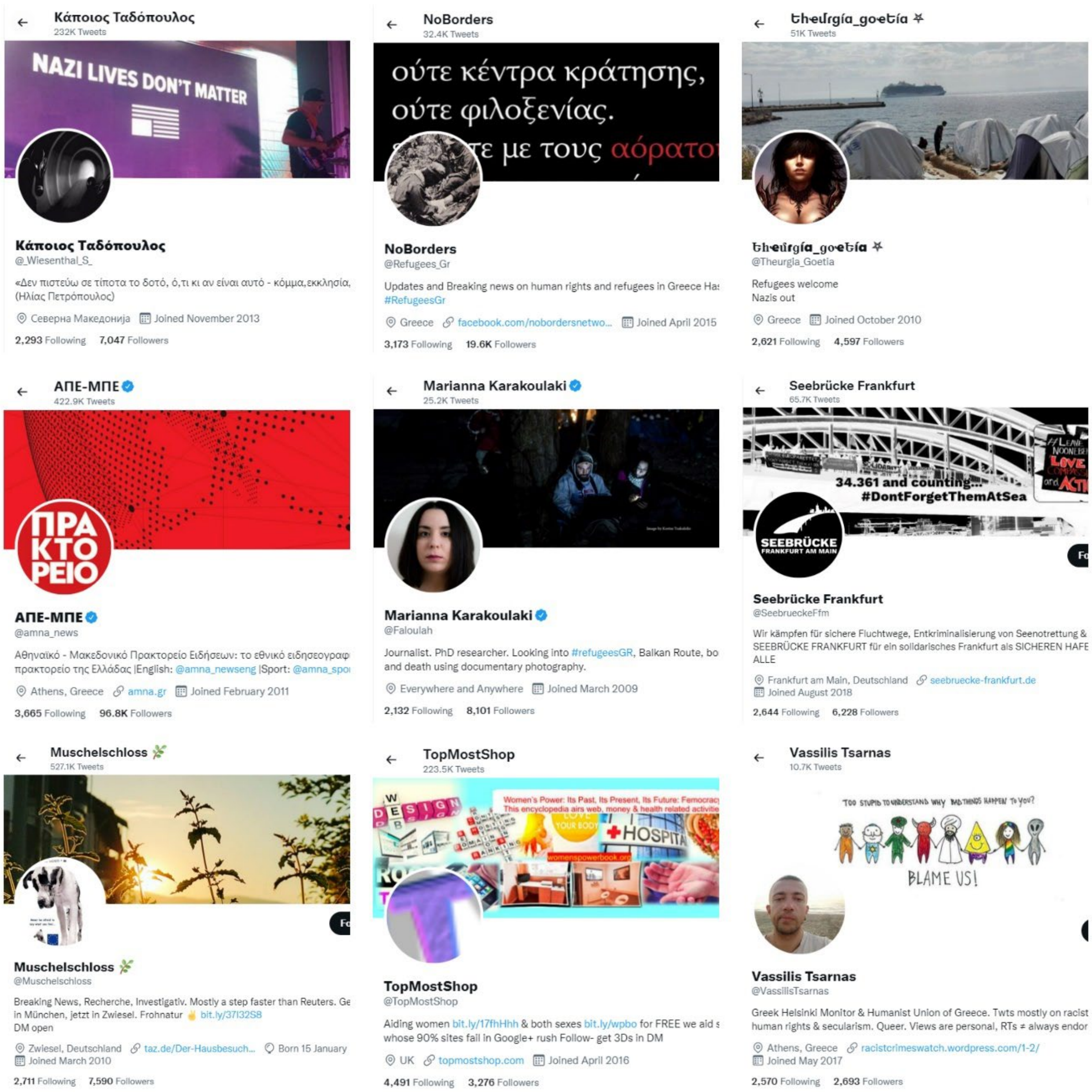

Figure 4-6: Most active Twitter users using the hashtag #RefugeesGR

Since there is a network analysis component in this dissertation, it will be useful to further investigate Twitter accounts posting about #RefugeesGR according to their verification status on the platform. To receive the "verification badge" on Twitter, a user account must be authentic, notable, and active[31], while this status is generally provided to accounts of high public interest. Figure 4-7 below showcases the classification of user accounts within the

[31] help.twitter.com/en/managing-your-account/about-twitter-verified-accounts

dataset based on verified status, with 331 Twitter users being verified and 11,423 non-verified. The observed 3% of verified accounts within the dataset is significantly higher than the general percentage of verified accounts amongst Twitter users, which is less than 1%[32] (primary data is not published by Twitter and relevant information is provided through secondary sources). This might suggest that there is more general interest amongst public figures, government agencies, news organizations and brands in the hashtag #RefugeesGR.

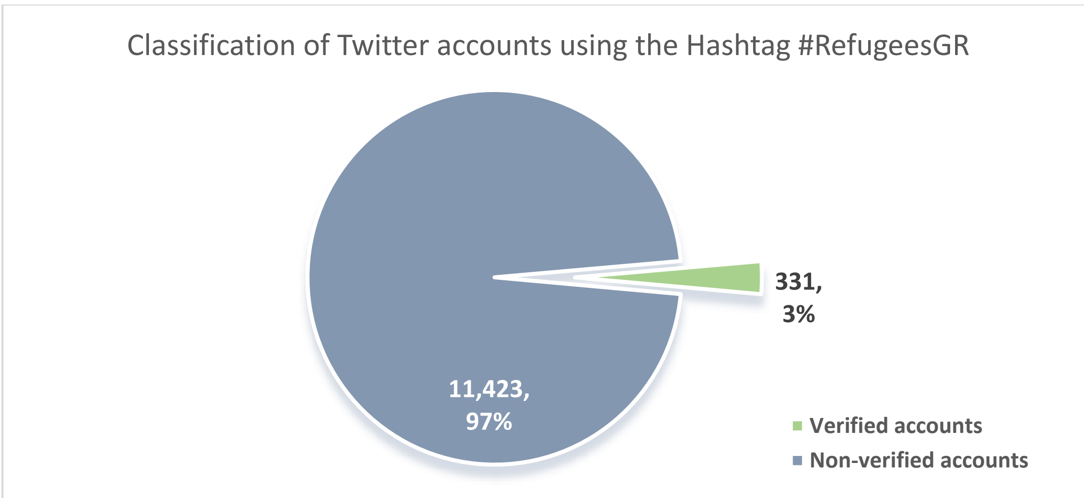

Figure 4-7: Classification of Twitter users by verified status

Looking further into these 331 verified Twitter accounts that use the hashtag #RefugeesGR, it is clear that the vast majority consists of news outlets and journalists, according to Figure 4-8 below, which depicts the most popular verified accounts (with more than 10,000 followers) within the dataset. Official government accounts and politicians, as well as international humanitarian agencies and humanitarian organizations also have significant presence amongst the verified users.

[32] www.vox.com/22444961/twitter-verification-process-verified-blue-checkmark-jack-dorsey

Figure 4-8: Verified Twitter user accounts by popularity

Similarly, Figure 4-9 depicts verified users based on the usage of the hashtag #RefugeesGR (at least 4 tweets per account) and, once more, news outlets and journalists comprise the majority of these accounts, with politicians, government bodies and humanitarian organizations following closely.

Figure 4-9: Verified Twitter user accounts by tweet volume that includes the hashtag #RefugeesGR

Further categorization of the Twitter accounts and the interactions among them will be the subject of discussion in Chapter 4.5.

4.2 Comparative yearly analysis

This part of the analysis will focus on what the Twitter users who used the hashtag #RefugeesGR were concerned about on a broad level. For each year within the period 2015 to 2021, the volume of the tweets will be analyzed and the events around the dates of the highest volume of tweets will be investigated further, in order to understand what the user activity revolved around throughout the years. Following the event analysis, the most popular hashtags per year will be presented, along with the most commonly used words per year, in order to investigate if there is common ground to be found and, lastly, the most popular tweets per year will be presented. This analysis will lay the ground for the topic modelling findings, presented in Chapter 4.4.

It is important to note that the events presented in this chapter do not necessarily include all the major events around the refugee crisis in Greece over the 2015 – 2021 period, but rather the events that sparked the most conversation on Twitter amongst users who utilized the hashtag #RefugeesGR.

4.2.1 Year 2015

According to UNHCR, 2015 was the year when the number of new arrivals in Europe through the Mediterranean route peaked, with more than one million people making the journey and requesting asylum in EU territory, corresponding to approximately a 360% increase compared to the previous year, 2014[33]. The vast majority of those new arrivals passed through Greece, with more than 860,000 people in total and some 800 dead and missing persons[34]. Daily arrivals during 2015 started surpassing the threshold of 1,000 persons per day in June, while the period of August to December had more than 100,000 persons arriving per month.

[33] data.unhcr.org/en/situations/mediterranean
[34] data.unhcr.org/en/situations/mediterranean/location/5179

The number of daily tweets revolving around #RefugeesGR started regularly surpassing 100 tweets per day after July 2015, while there are six distinct peaks of daily tweet volume, according to Figure 4-10 below, which will be analyzed further.

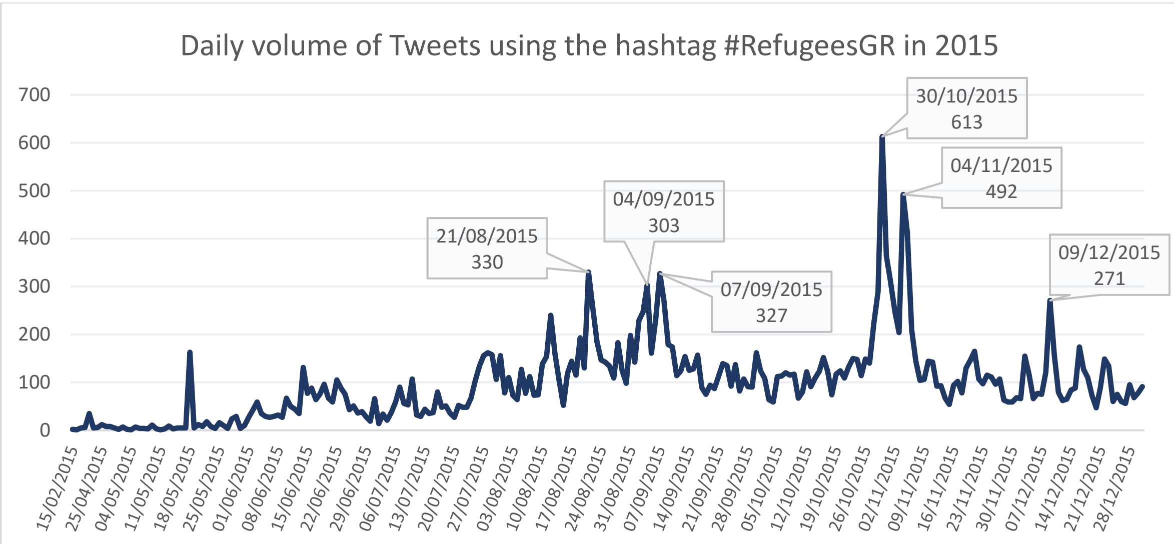

Figure 4-10: Daily volume of Tweets and peak dates in 2015

In early 2015, as the influx of new arrivals steadily increased in Greece, more and more people made the journey to cross the borders from Greece into the Western Balkans on foot, with the intention to travel to other European countries. Frontex recorded a 663% increase[35] of border crossings from Greece during the first six months of 2015 (compared to the same period in 2014) and there was little border resistance in the process. On 21 August 2015, North Macedonia declares a state of emergency in the country due to the high numbers of people on the move, effectively closing its southern border. As the norm up to that point was for about 2,000 persons crossing from Greece to North Macedonia on a daily basis, hundreds of people continue attempting to cross and clashes with North Macedonian border police[36] take place, which peaks Twitter user's attention. On 4 September 2015, Vice-President of the EU Commission, Frans Timmermans, and EU Migration and Home Affairs Commissioner, Dimitris Avramopoulos, visit Kos island to discuss solutions for the humane accommodation

[35] frontex.europa.eu/assets/Publications/Risk_Analysis/WB_Q2_2015_report.pdf
[36] www.dw.com/en/macedonia-disperses-migrants-with-tear-gas-stun-grenades/a-18663901 and www.bbc.com/news/world-europe-34014353

and processing of asylum-seekers in Greece, while Amnesty International publishes a report on a violent attack towards asylum-seekers, activists, and humanitarian workers on the island[37]. On 7 September 2015, half of the tweets in the dataset mention Lesvos and the dire situation on the island, where more than 15,000 asylum-seekers await registration in Mytilene[38]. During the last days of October 2015, there are 5,500 daily arrivals on average on the Greek islands, according to a video[39] by Médecins du Monde Greece circulating on Twitter. During the same period, the news of three shipwrecks[40] on Lesvos, Kalymnos and Rhodes with multiple casualties are also among the top discussed topics. On 4 November 2015, European Parliament President, Martin Schulz, and Greece's Prime Minister at the time, Alexis Tsipras, visit Lesvos and the first hotspot in Greece, Moria[41][42], with many tweets urging both to take action. On 9 December 2015, a large Greek police operation evacuates most people from the camp of Idomeni and gets the attention of Twitter as journalists in the area post about police obstructing their work[43].

Moving on from the events around the peak dates of tweet volume, the most popular hashtags used in conjunction with #RefugeesGR during 2015 are mainly about places where asylum-seekers are gathered, namely #Lesvos, #Kos, #Idomeni, #Chios and #PedionAreos, while solidarity hashtags, such as #RefugeesWelcome, #SafePassage, #OpenEuBorders are also quite popular during that time. Figure 4-11 below provides a visual representation of the most popular hashtags used in conjunction with #RefugeesGR during 2015.

[37] http://www.amnesty.org/en/latest/press-release/2015/09/greece-refugees-attacked-and-in-hellish-conditions-on-kos
[38] www.theguardian.com/world/2015/sep/07/lesbos-on-verge-of-explosion-as-refugees-crowd-greek-island
[39] www.youtube.com/watch?v=fRIGEyOPYBc
[40] www.tovima.gr/2015/10/30/society/tria-nea-nayagia-sto-aigaio-23-nekroi-anamesa-toys-13-paidia
[41] www.topontiki.gr/2015/11/05/tsipras-ke-soults-ichan-tin-atichia-na-doun-me-ta-matia-tous-mia-varka-na-erchete-gemati-prosfiges-photos-video
[42] www.reuters.com/article/us-europe-migrants-greece-idUSKCN0SU28420151105
[43] observers.france24.com/en/20151209-greek-police-macedonia-migrant-idomeni-camp-photos and www.keeptalkinggreece.com/2015/12/09/greek-police-operation-evacuates-idomeni-from-economic-migrants

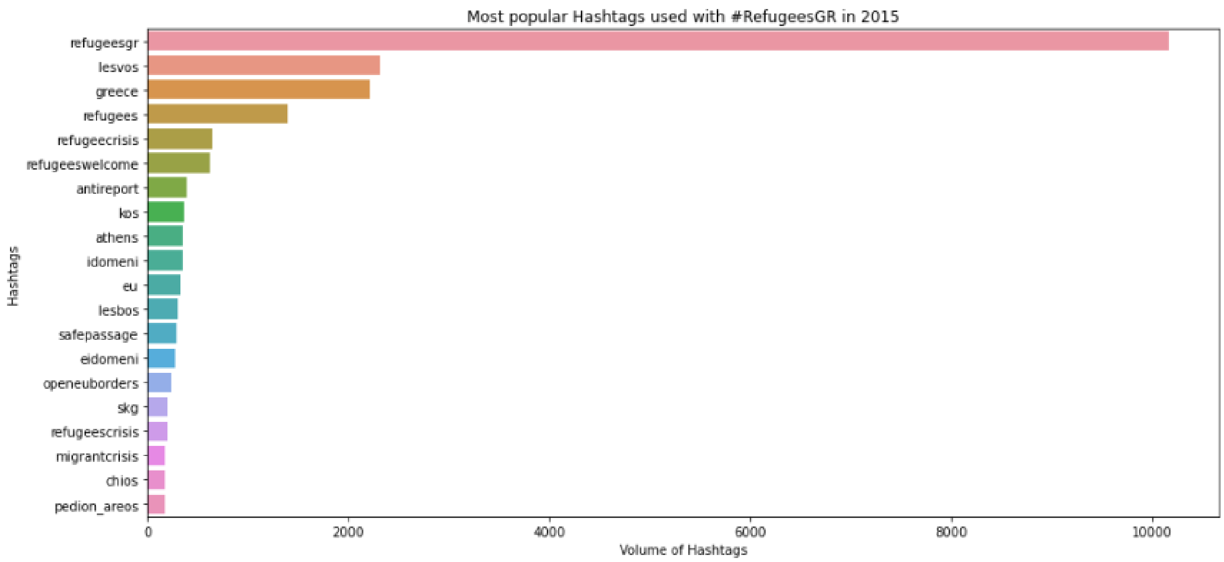

Figure 4-11: Most popular hashtags used alongside #RefugeesGR in 2015

Similarly, there are many solidarity and call-to-action words amongst the most common ones found in 2015 tweets, such as: "help", "volunteer", "need", "solidarity" and "support". The most popular location-based words are "island" and "border", while the only nationality appearing within the top words is "Syrian". The words "police", "migrant" and "camp" are also quite prominent in the 2015 dataset, with about 400 appearances each. Figure 4-12 provides an insight of the most used words in tweets regarding #RefugeesGR during 2015.

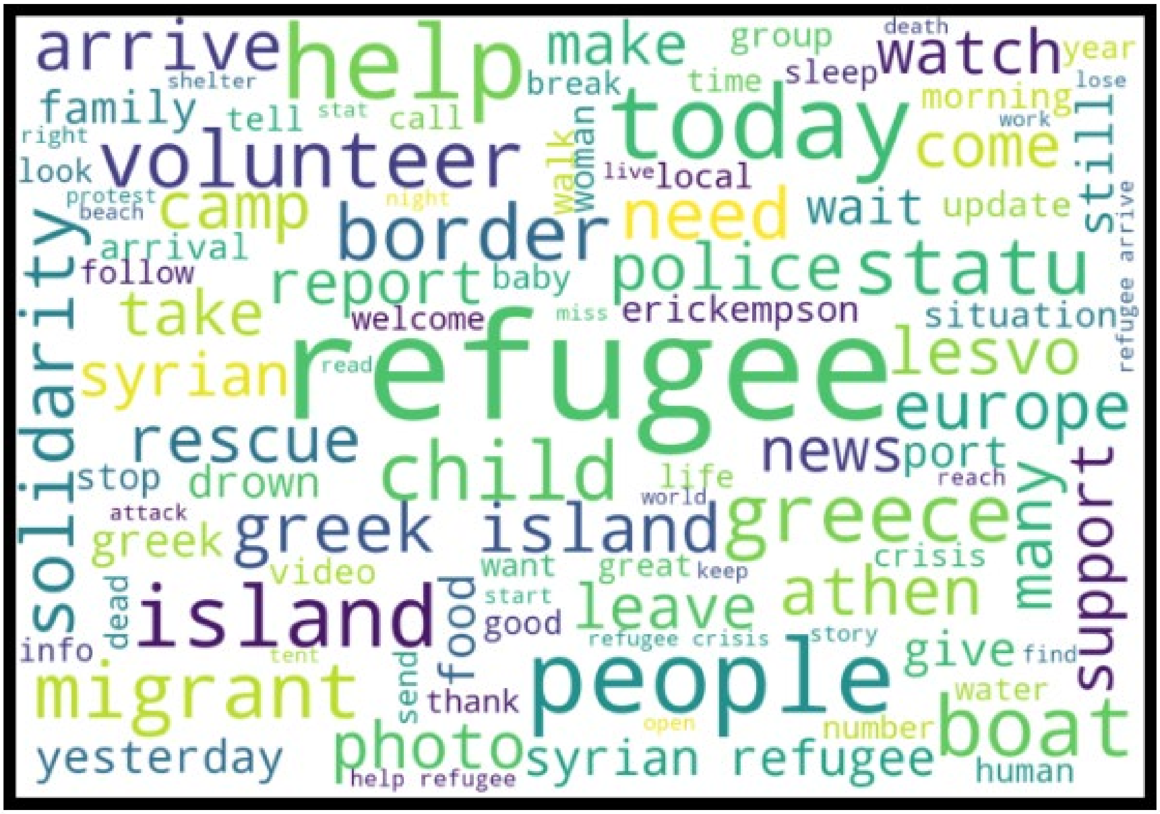

Figure 4-12: Top 100 most used words in tweets with the hashtag #RefugeesGR in 2015

To conclude the year, Figure 4-13 below showcases the most popular tweets on #RefugeesGR during 2015, based on the number of likes received.

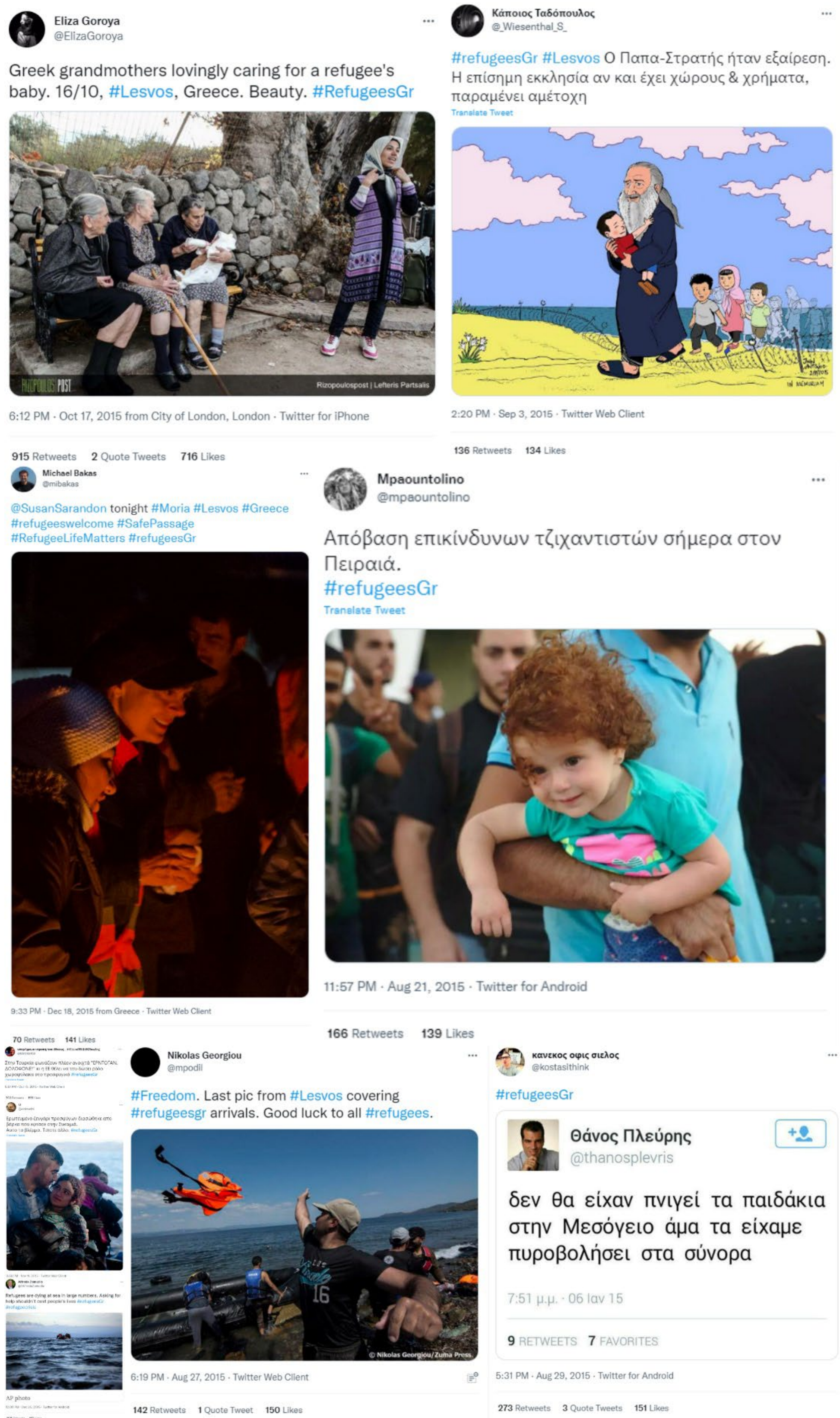

Figure 4-13: Most liked tweets in 2015

4.2.2 Year 2016

Arrivals in Greece during 2016 decreased by 80% compared to 2015, with more than 175,000 persons arriving in total and some 440 dead and missing persons[44]. Most of the arrivals during 2016 took place during the first three months of the year.

The arrivals trend in 2016 is on par with the number of daily tweets of the same year, with most of the activity in Twitter around the hashtag #RefugeesGR observed during the period January to March, according to Figure 4-14 below, where six peaks are marked for further discussion.

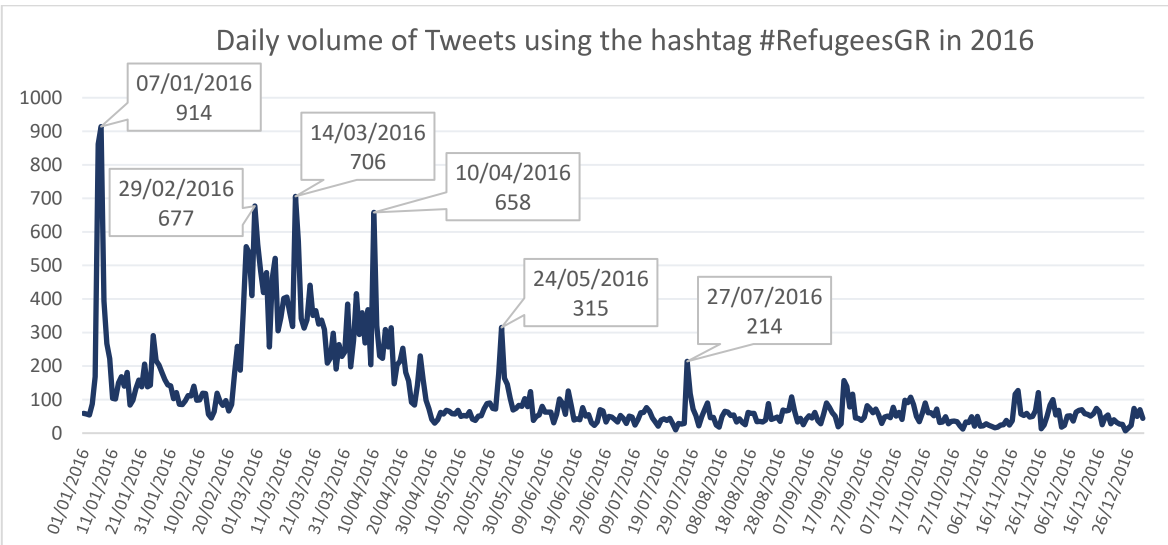

Figure 4-14: Daily volume of Tweets and peak dates in 2016

On early January 2016, more than 1,500 tweets mention an Avaaz petition[45] asking citizens to sign in order for the Nobel Peace Prize to be awarded to the Greek islanders, who have shown solidarity to asylum-seekers arriving in Greece. The situation in Idomeni camp, at the border between Greece and North Macedonia, continues to escalate during February and March 2016, as North Macedonian authorities only allow small numbers of people to enter the country per day (compared to thousands a few months prior). As asylum-seekers continue to travel to Idomeni in the hopes of being allowed into North Macedonia, many not knowing

[44] data.unhcr.org/en/situations/mediterranean/location/5179
[45] secure.avaaz.org/campaign/el/nobel_to_greek_islanders/?wfJsckb

that the borders are not effectively open, a serious bottleneck of 14,000 people (at its peak) stranded in the area is forming[46]. On 14 March 2016, more than a thousand asylum-seekers leave Idomeni in an attempt to reach North Macedonia through an alternative route, crossing the swelled river[47] at the border by forming a human chain. Three asylum-seekers drown during the river crossing attempt[48] and the event is heavily reflected in daily tweet volume. Amidst this growing pressure at the border, the tweets during that period also mention the situation that is developing at Piraeus port, where more than 5,000 asylum-seekers are also residing in a makeshift camp[49], as boats from the islands arrive, but people are unwilling to be transferred to official government camps out of fear of being stranded in Greece and in hope that the borders will re-open. In addition, as the EU-Turkey Statement[50] takes effect on 18 March 2016, tweets are also focused on what that practically means for newly arrived asylum-seekers in Greece after that date. The EU-Turkey Statement essentially establishes two different international protection procedures, depending on the time an asylum-seeker arrived in Greece, thus creating two sub-populations facing different realities and needs. For people arriving after 18 March 2016, applying for asylum in Greece becomes their only option. The escalation of the situation in Idomeni continues, despite the efforts of the Greek government to completely evacuate the site, which also results in the formation of a second unofficial camp in the same geographical area, in the parking lot of a gas station rest-stop just outside the town of Polykastro[51]. On 10 April 2016, serious clashes between asylum-seekers and police at the border between Greece and North Macedonia take place, with North Macedonian police using tear gas, rubber bullets and water cannons in an effort to push people back from the border area[52]. The event caused the Greek government and International Humanitarian Organizations and Agencies to issue condemning statements with reporters and volunteers tweeting live from the area. On 24 May, a large police operation moves to evacuate the remaining 8,000 asylum-seekers from Idomeni to newly established camps

[46] www.amnesty.org/en/latest/press-release/2016/02/idomeni-border-crisis and www.theguardian.com/world/2016/mar/02/idomeni-greece-refugee-march-abruptly-cut-short
[47] www.youtube.com/watch?v=rhKFKGV23H8
[48] www.reuters.com/article/uk-europe-migrants-macedonia-idUKKCN0WG0ZS
[49] www.hrw.org/news/2016/03/24/greece-humanitarian-crisis-athens-port
[50] www.consilium.europa.eu/en/press/press-releases/2016/03/18/eu-turkey-statement
[51] deeply.thenewhumanitarian.org/refugees/articles/2016/06/14/photo-essay-eko-camp-before-evacuation
[52] ecre.org/tear-gas-and-rubber-bullets-for-refugees-in-idomeni

across Greece[53], with International Organizations and Agencies supporting the non-use of force, albeit raising concerns over the facilities in the camps[54], which is reflected in tweets during that time. The tweet volume on #RefugeesGR deflates after these events on early 2016, however the simultaneous police raids and evacuations of squats in Thessaloniki, Athens and the informal camp at the port of Piraeus[55], where asylum-seekers were residing, peaks again the interest of Twitter users on 27 July.

Having explored the events that sparked Twitter conversation during 2016, taking a closer look at the most popular hashtags used alongside #RefugeesGR during that time is affirming the previous findings. Similar to 2015, location-based hashtags continue to be very popular, although #Idomeni is much more prominent in 2016 than it was in 2015, even surpassing #Lesvos. Also, #Piraeus and #Moria are among the most popular hashtags this year, while #Kos is not, which is on par with the event analysis above. Solidarity hashtags, such as #RefugeesWelcome and #SafePassage continue to be widely used in tweets and the hashtag #EuTurkeyDeal is making its first appearance this year. Figure 4-15 below lists the twenty most popular hashtags on #RefugeesGR in 2016.

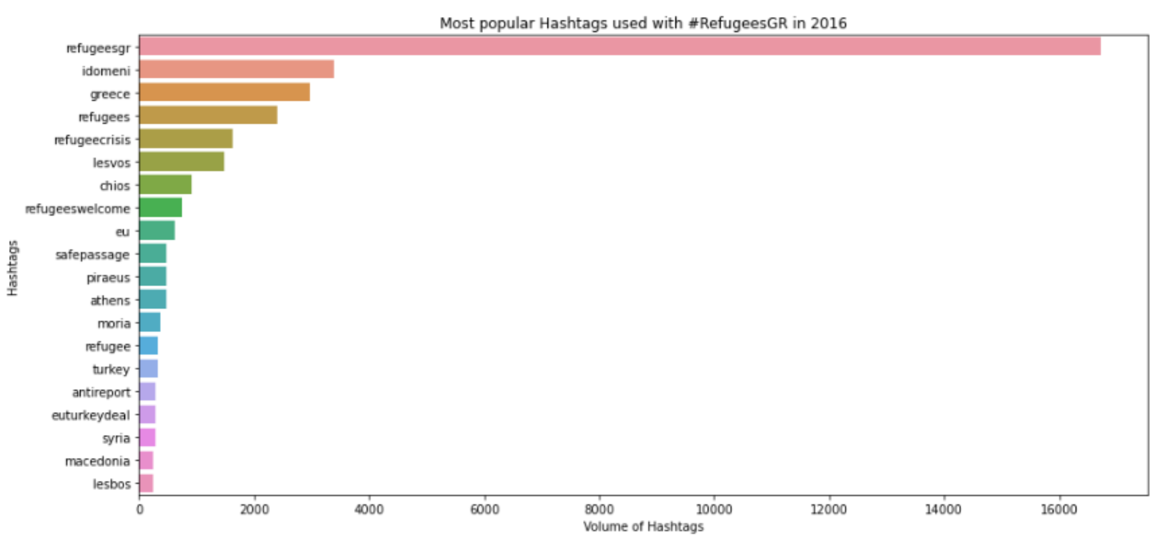

Figure 4-15: Most popular hashtags used alongside #RefugeesGR in 2016

53 www.bbc.com/news/world-europe-36358891
54 www.aljazeera.com/news/2016/5/25/greek-police-move-to-shut-down-idomeni-refugee-camp
55 ceasefiremagazine.co.uk/solidarity-criminalised-anger-greek-police-raids-refugee-housing-squats-camps

While in 2015, the word “help” was the most prominent one amongst tweets on #RefugeesGR, in 2016 the most commonly used word was “camp”. Location based words, such as “border” and “island” continue to be among the most used ones in 2016 tweets, however the order has changed significantly, as “border” comes second in use, while “island” nineteenth. Call-to-action words continue to be popular in 2016, much like in 2015. The word “police” gains popularity this year, while the word “protest” makes an appearance in the twenty most used words in 2016 tweets. Figure 4-16 below contains the most used words in tweets about #RefugeesGR in 2016.

Figure 4-16: Top 100 most used words in tweets with hashtag #RefugeesGR in 2016

Wrapping up the findings for this year, Figure 4-17 below depicts the most popular tweets about #RefugeesGR during 2016, according to the number of likes each tweet received.

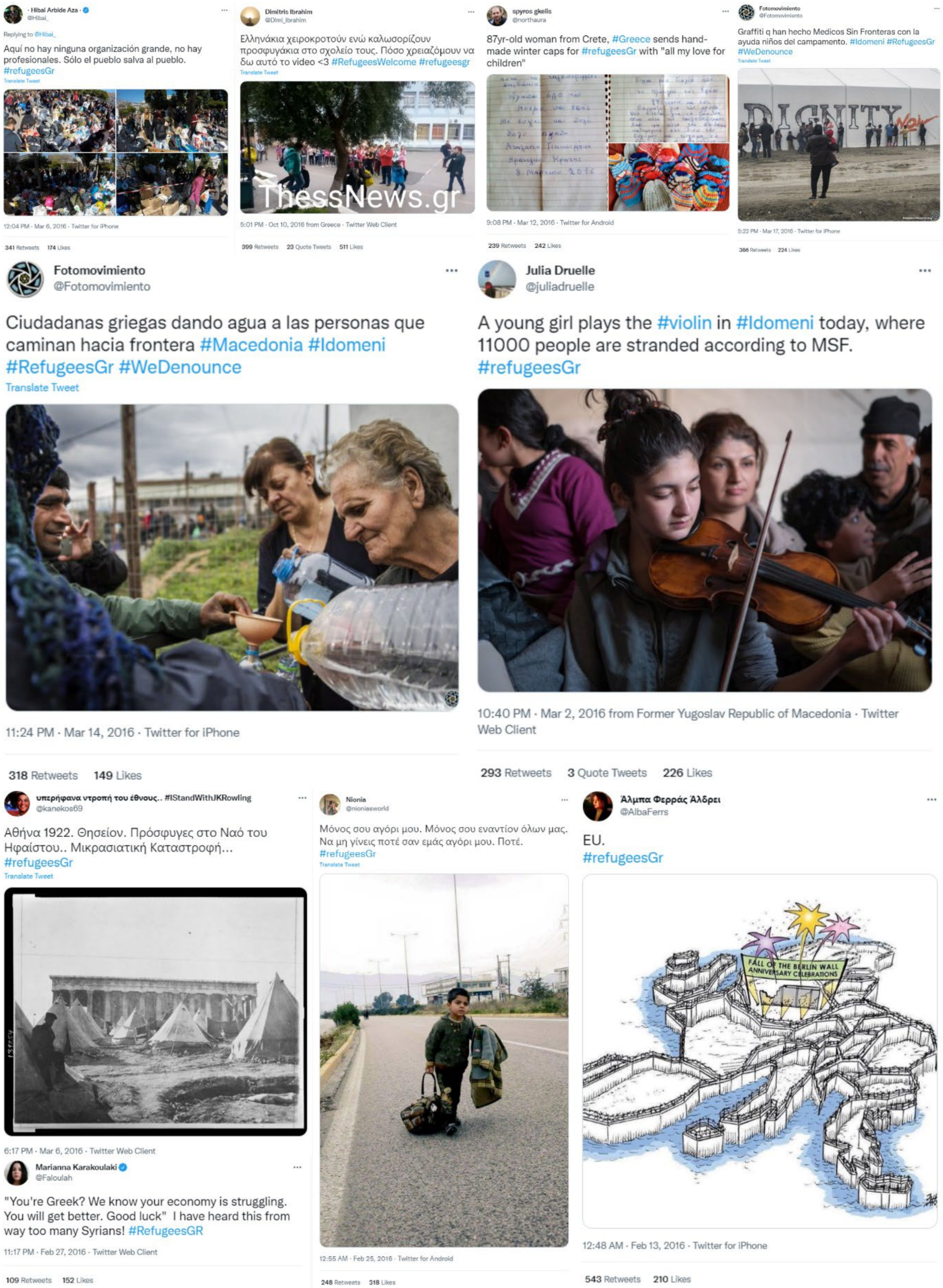

Figure 4-17: Most liked tweets in 2016

4.2.3 Year 2017

Arrivals in Greece during 2017 further decreased by 80% compared to 2016, with some 36,000 arrivals in total and about 60 dead and missing persons[56] according to UNHCR data. The period with the most arrivals during the year was August to November, with more than 3,000 arrivals per month.

The number of tweets also significantly decreased in 2017, by more than 70% compared to 2016. There are only four peaks marked on Figure 4-18, for having more than 100 tweets per day.

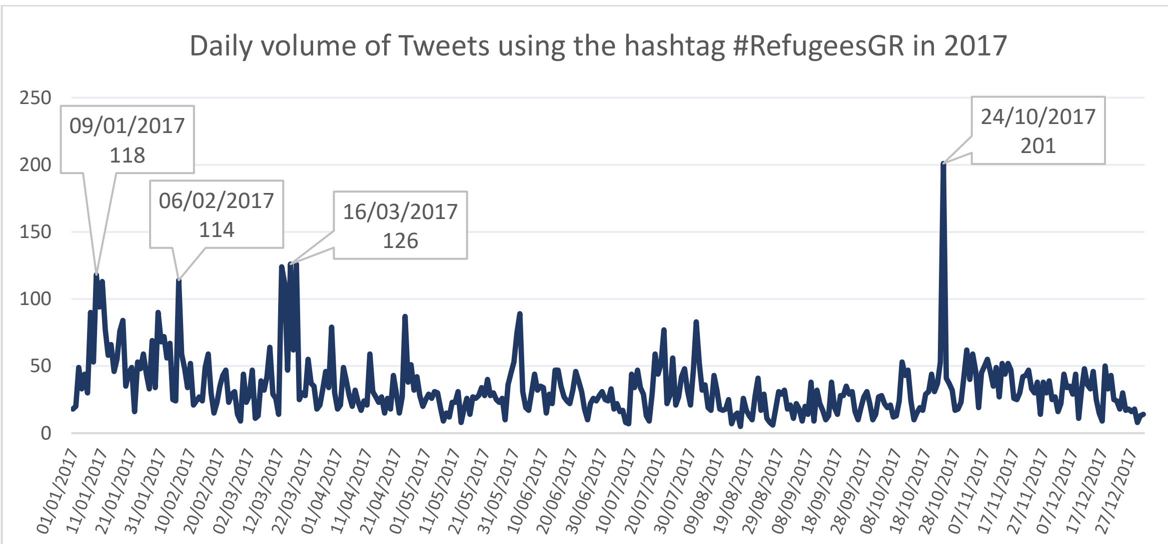

Figure 4-18: Daily volume of Tweets and peak dates in 2017

While the scale of tweet volume on #RefugeesGR is much lower in 2017 compared to 2016, events in late January regarding the harsh weather conditions[57] in camps across Greece and the deaths of three asylum-seekers[58] sparked interest. On 6 February 2017, the (at-the-time) Greek Minister for Migration Policy, Yiannis Mouzalas, was not allowed[59] to enter the facilities at Elliniko, the former airport in Athens, where the living conditions for asylum-

[56] data.unhcr.org/en/situations/mediterranean/location/5179
[57] www.nytimes.com/2017/01/11/world/europe/greece-refugees-crisis-winter-storms.html
[58] www.reuters.com/article/europe-migrants-greece-idINKBN15E1G6
[59] gr.euronews.com/2017/02/06/mouzalas-elliniko-visit-refugees

seekers were reported to be below humanitarian standards. The protest at Elliniko[60] peaked tweet volume with users sharing pictures from the Minister's visit. On mid-March 2017, two separate events sparked Twitter's interest. Two police raids and evictions from refugee squats on 13 March[61] caused protests in Athens, as many asylum-seekers were detained in the process. On 16 March, the International Rescue Committee (IRC), the Norwegian Refugee Council (NRC) and Oxfam published a report on "how the EU-Turkey Statement has turned Greece into a testing ground for European Union policies that are eroding the rights of refugees and asylum seekers"[62], with tweets mostly revolving around its findings. With arrivals on Greek islands being high in the second half of the year, conditions in Moria camp in Lesvos and VIAL camp in Chios continue to deteriorate. As a result, tensions and protests on both islands escalate[63], with tweets calling for action, especially regarding winterization of the camps.

Continuing with the most popular hashtags used in conjunction with #RefugeesGR during 2017, it's apparent that #Idomeni is no longer used, as the camp completely shut down in 2016. Twitter focus during this year has shifted towards the islands, with hashtags such as #Lesvos, #Moria, #Chios and #Samos being the most prominent. Hashtag #Kos continues not to be among the most popular hashtags for 2017, but a new hashtag #OpenTheIslands makes an appearance, referring to the geographical movement restrictions imposed to asylum-seekers on the islands during the year. Similar to 2016, #EuTurkeyDeal is also popular in 2017, as well as solidarity hashtags, such as #RefugeesWelcome and #SafePassage. The hashtag #HumanRights is also making its first appearance in the top twenty most used hashtags on #RefugeesGR and the whole visualization is portrayed in Figure 4-19.

[60] www.reuters.com/article/europe-migrants-greece-idUSL5N1FR26X
[61] enoughisenough14.org/2017/03/13/athens-khoras-social-centre-on-squat-evictions-refugeesgr-living-there
[62] www.nrc.no/resources/briefing-notes/how-eu-policies-are-eroding-protection-for-refugees-in-greece
[63] areyousyrious.medium.com/ays-daily-digest-20-10-2017-high-tension-on-greek-islands-continues-764af48ba44

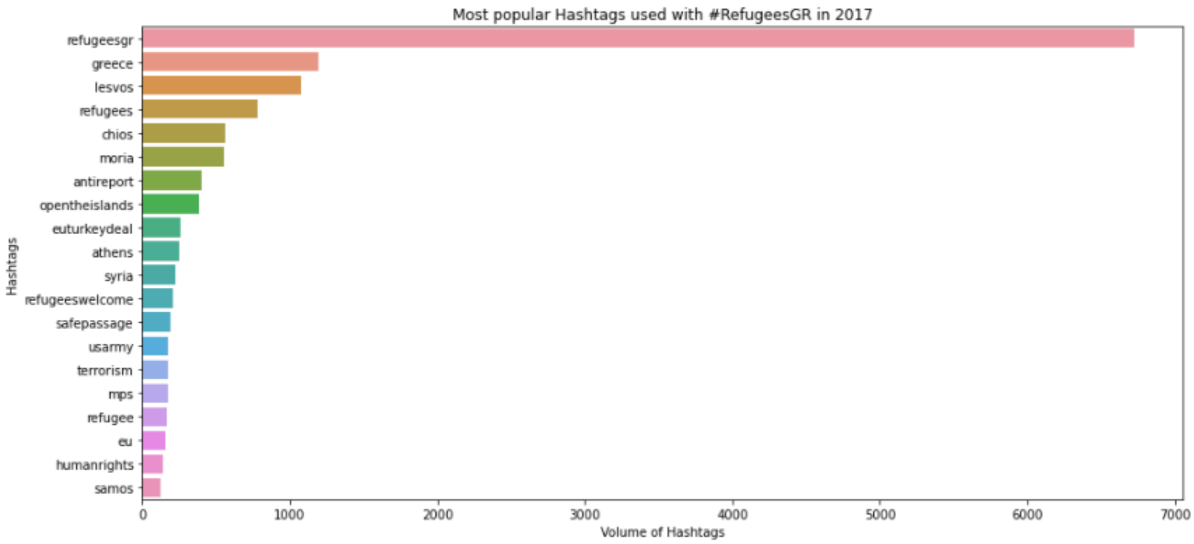

Figure 4-19: Most popular hashtags used alongside #RefugeesGR in 2017

Much similar to 2016, the word "camp" is the most used one in tweets about #RefugeesGR during 2017. However, location-based words only include "island" in the top twenty this year. Like previous years, solidarity and call-to-action words, such as: "help", "need", "support" and "solidarity" continue to be popular, but there are also distress words introduced in the top twenty, such as "strike", "hunger" and "condition", in conjunction with "protest", which was already popular in 2016. Figure 4-20 showcases the full picture of the hundred most commonly used words in tweets about #RefugeesGR during 2017.

Figure 4-20: Top 100 most used words in tweets with hashtag #RefugeesGR in 2017

To complete the summary findings for the year, Figure 4-21 below provides a collage of the most popular tweets on #RefugeesGR during 2017, based on the number of likes the tweets received.

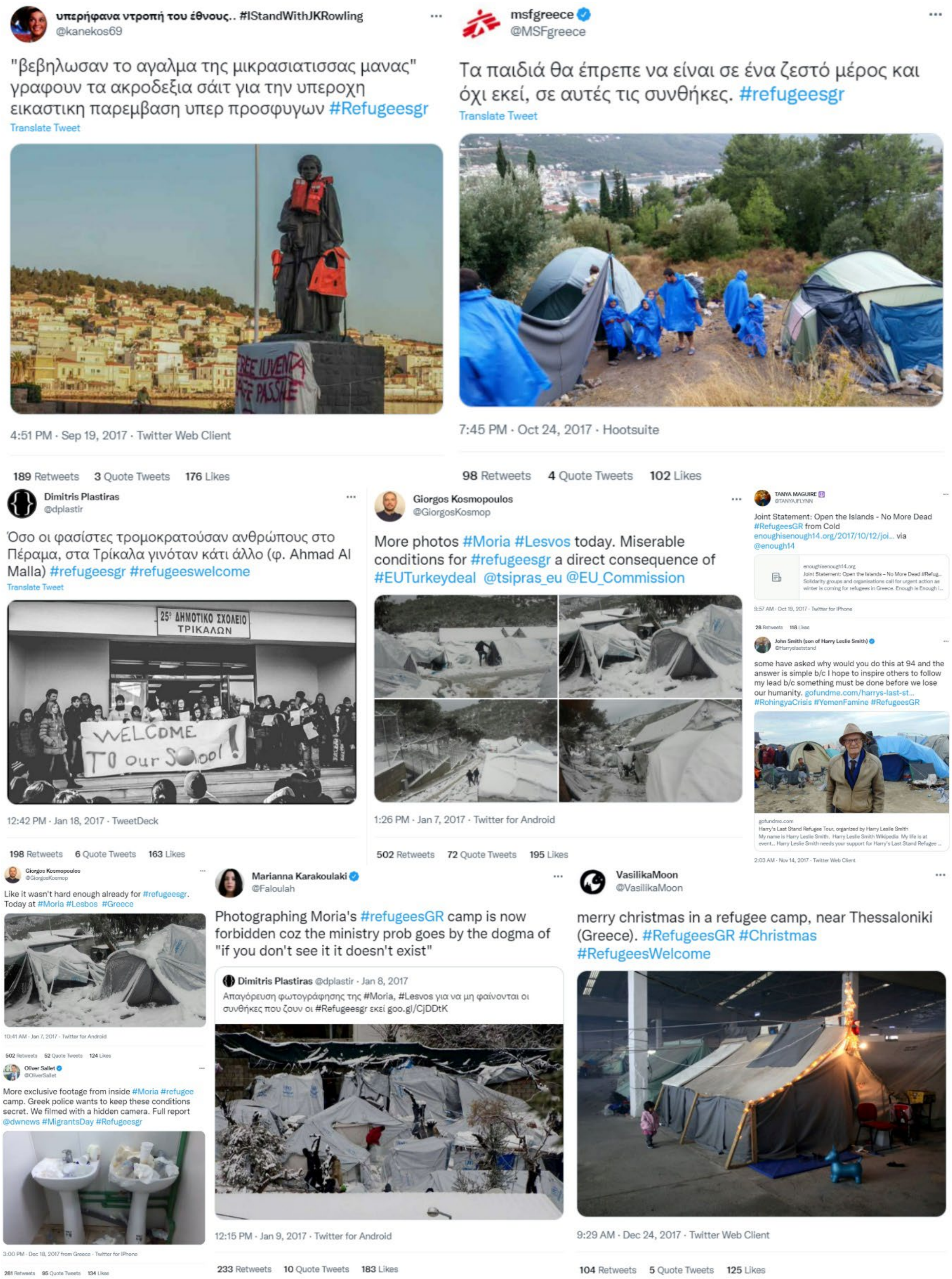

Figure 4-21: Most liked tweets in 2017

4.2.4 Year 2018

In 2018, Greece saw a 40% increase on new arrivals compared to 2017, with about 50,000 people arriving in total and some 170 dead and missing persons[64]. The arrivals trend per month was steadily above 3,000 people, with the majority recorded in April, September and October.

Despite the significant increase on arrivals, the number of tweets about #RefugeesGR in 2018 decreased by 50% compared to 2017. Thus, there are only three peaks that can be clearly distinguished among the daily volume of tweets in Figure 4-22, which will be analyzed further.

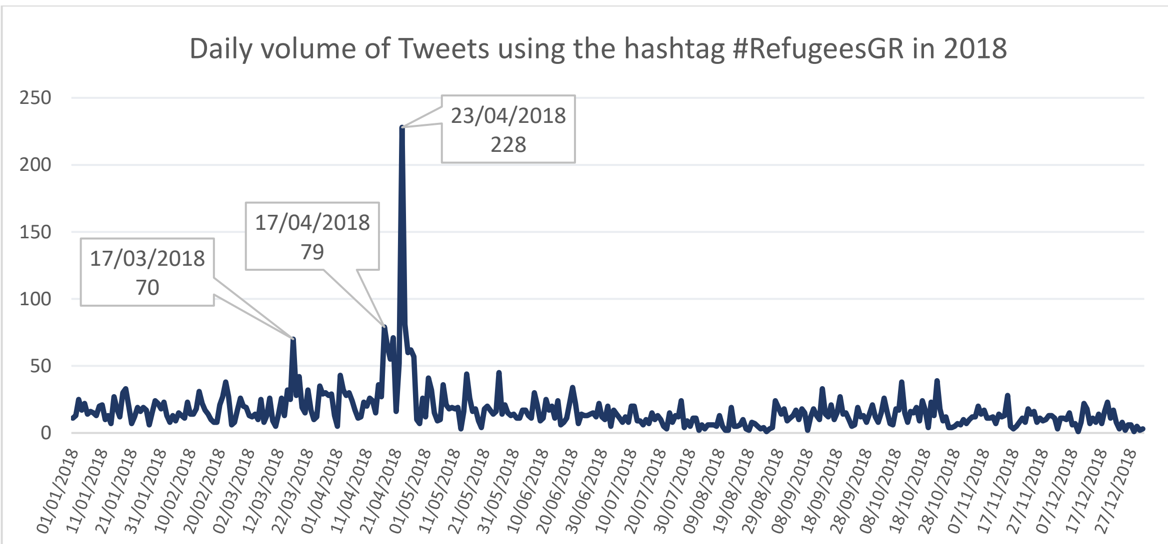

Figure 4-22: Daily volume of Tweets and peak dates in 2018

The first event that sparked Twitter interest in 2018 was the shipwreck of 17 March in Agathonisi, Samos, that resulted in 16 deaths[65]. On 17 April, Greece's Supreme Administrative Court issued a decision against movement restrictions imposed to asylum-seekers on the six Aegean islands, due to an appeal by the Greek Council of Refugees[66] and many tweets shared the news. However, the Greek government tabled a bill shortly after,

[64] data.unhcr.org/en/situations/mediterranean/location/5179
[65] apnews.com/article/98bd99cbc9ef4e56adad3a1ee2e823a9
[66] www.kathimerini.gr/society/959492/nea-dedomena-gia-toys-prosfyges-me-apofasi-toy-ste

allowing Greek authorities to restrict the freedom of movement of international protection applicants to a specific part of the Greek territory[67]. On the night of 22 April, a group of some 200 far-right extremists attacked asylum-seekers who were peacefully protesting in Sapphous square in Mytilene, Lesvos, injuring a dozen people[68]. The clashes lasted through the night and more than half of the tweets on 22 and 23 April were discussing these events and subsequent arrests. In addition to those distinct events, more than 250 tweets in April 2018 mention the trial of 35 asylum-seekers, who were arrested in Moria camp in July 2017 following a protest and a relevant hashtag linked to this case became popular[69] during that time.

Moving on from the events that sparked Twitter's interest in 2018 to exploring the top-twenty most popular hashtags of the year, location-based hashtags only include #Lesvos, #Moria and #Chios, while #OpenTheIslands is now the fifth most popular one (excluding #RefugeesGR). The hashtag #EuTurkeyDeal remains of interest among Twitter users and so do solidarity-based hashtags, such as #RefugeesWelcome, #HumanRights and #FreeTheMoria35, which is further explained in the event analysis above in connection to a trial involving asylum-seekers. Figure 4-23 provides a complete list of the most used hashtags during 2018, in conjunction with #RefugeesGR.

[67] reliefweb.int/report/greece/greece-note-legal-changes-proposed-greek-government-19-april-2018-greece-s-reception

[68] www.aljazeera.com/news/2018/4/23/far-right-attacks-increase-tension-in-greeces-lesbos and www.dw.com/en/lesbos-far-right-attack-on-migrants-leaves-several-injured/a-43499091

[69] freethemoria35.wordpress.com

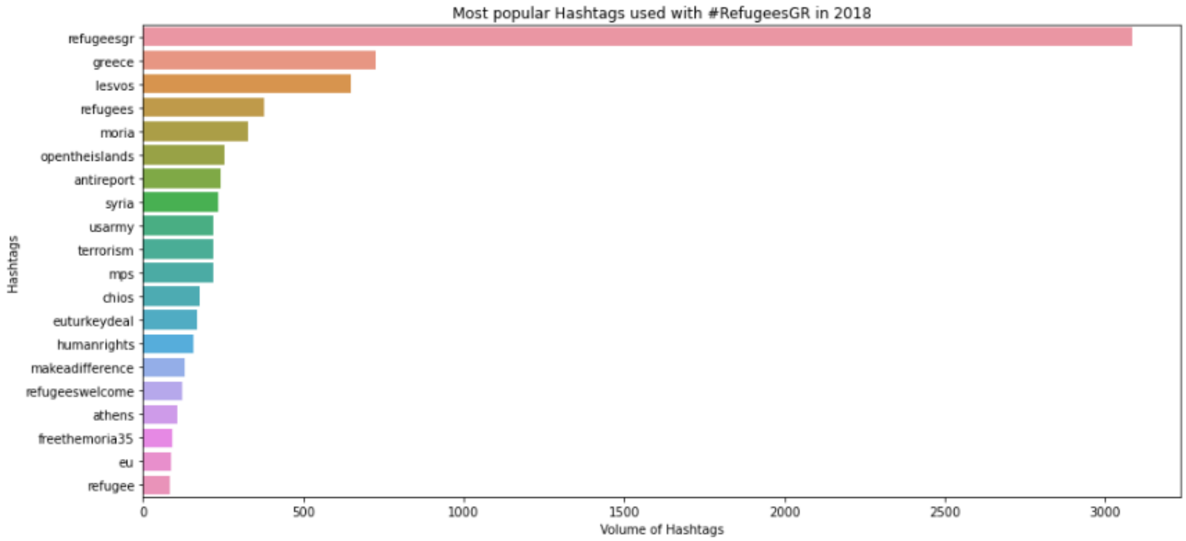

Figure 4-23: Most popular hashtags used alongside #RefugeesGR in 2018

Similar to 2017, the words "people" and "camp" are the two most common words in tweets regarding #RefugeesGR in 2018. Location-based words continue to include "island" in the top twenty this year, while the word "moria" is also making an appearance in that list. Solidarity and call-to-action words seem to be less popular in 2018, in comparison with previous years, with only "help" and "support" included in the top twenty this year. As in the 2017 tweets, distress words continue to be popular in 2018, with "strike" and "condition", as well as "burn" for the first time in this list. The full one hundred most used words in tweets about #RefugeesGR are shown in Figure 4-24 below.

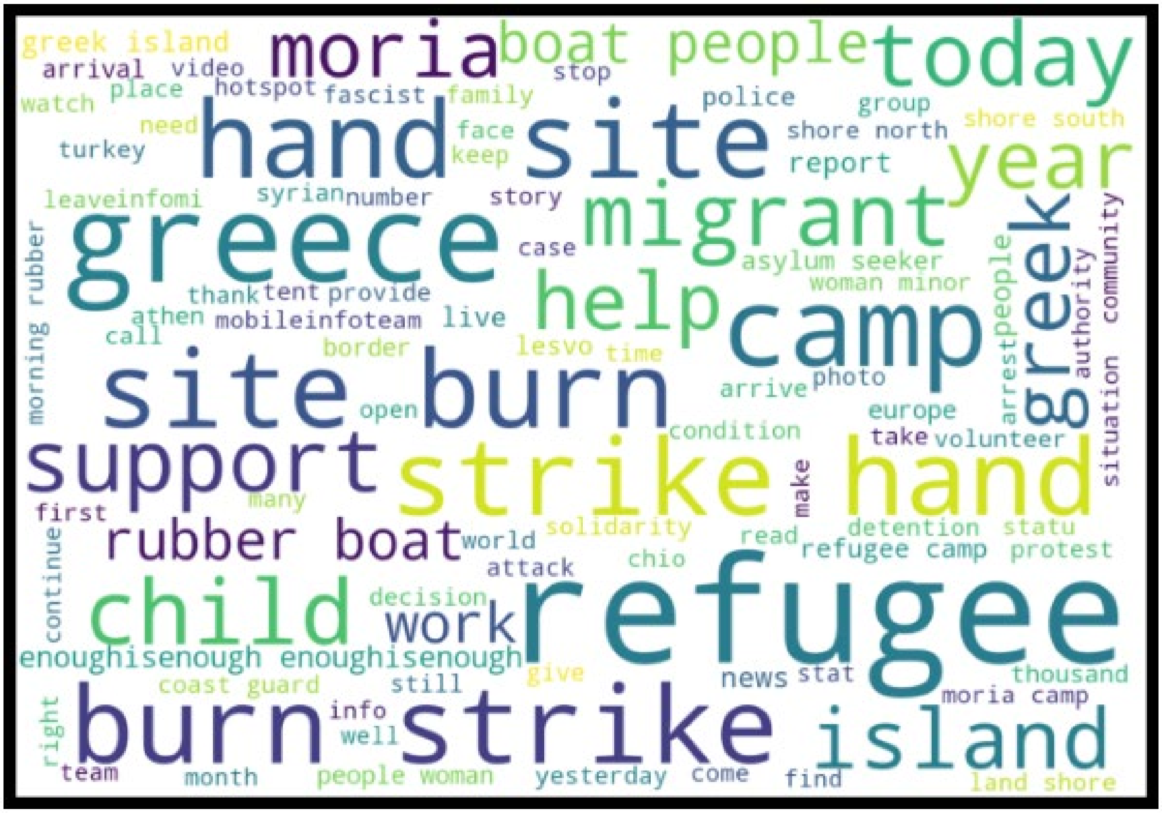

Figure 4-24: Top 100 most used words in tweets with hashtag #RefugeesGR in 2018

Finalizing the findings for the year, Figure 4-25 below showcases the most popular tweets on #RefugeesGR during 2018, according to the number of likes each tweet received.

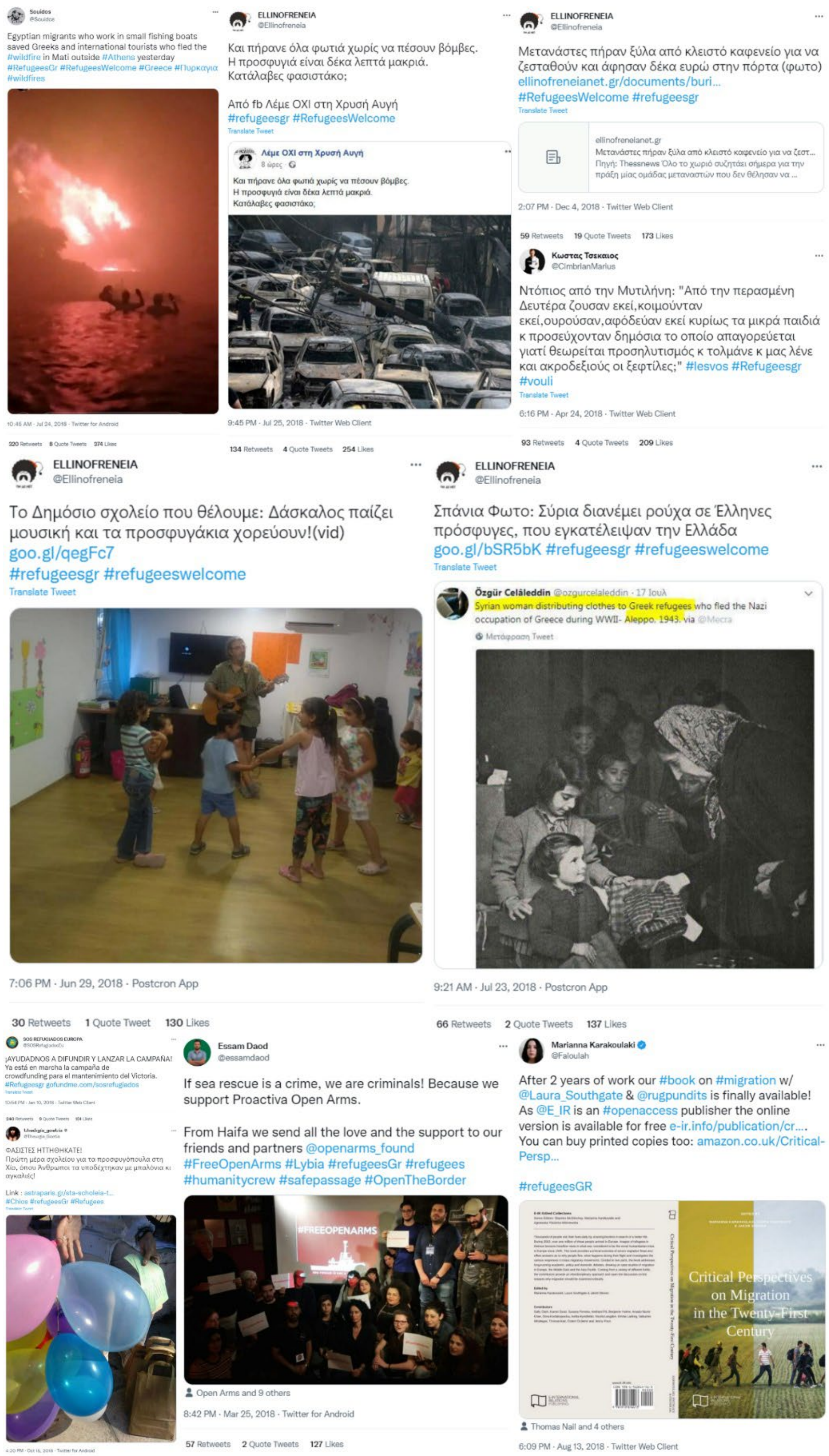

Figure 4-25:Most liked tweets in 2018

4.2.5 Year 2019

Arrivals to Greece in 2019 further increased by almost 50% compared to 2018, with more than 74,000 people arriving in total and about 70 dead and missing persons[70]. The proportion of persons dead and missing at sea in 2019 corresponds to 0.1% of the total arrivals, almost as low as 2015. There is a distinct trend of higher arrivals during the second half of 2019, with more than 5,000 arrivals per month during that period.

The number of tweets regarding #RefugeesGR in 2019 remained steady compared to 2018, despite the increase on arrivals, with a total of some 6,500 tweets. There are three peaks that reach above or near 100 tweets per day in Figure 4-26, which will be explored further.

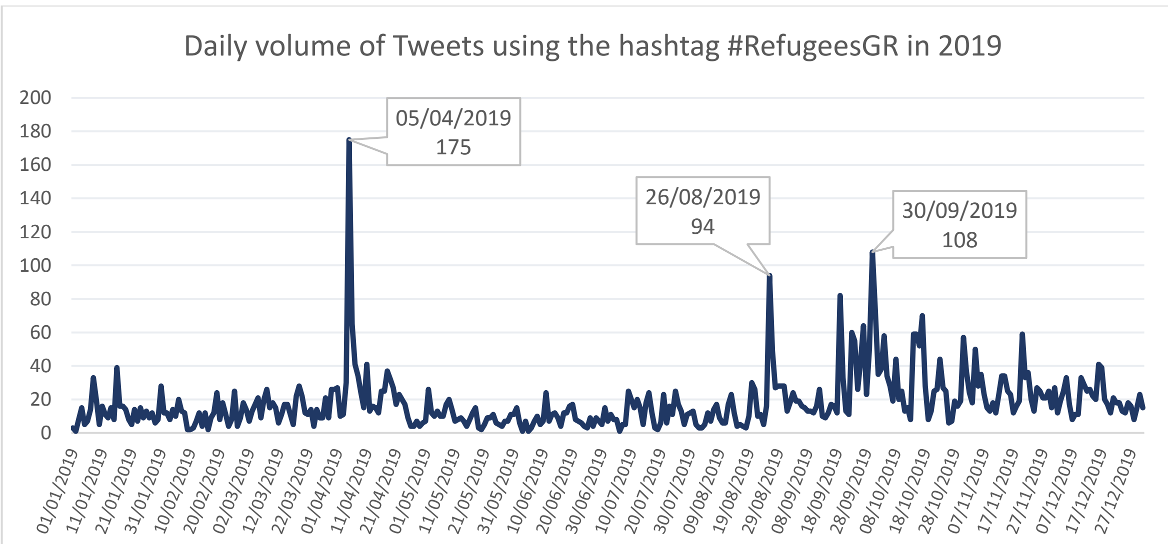

Figure 4-26: Daily volume of Tweets and peak dates in 2019

On early April 2019, rumors sparked among the refugee community that the borders to North Macedonia would re-open and, as a result, some 500 asylum-seekers and refugees gathered outside Diavata camp in Thessaloniki, in hopes of walking to the border area of Idomeni. Police tried to disperse the crowd, resulting in two days of clashes with the asylum-seekers[71], which were heavily tweeted. On 26 August 2019, police raids four squats in the Exarcheia

---

[70] data.unhcr.org/en/situations/mediterranean/location/5179
[71] balkaninsight.com/2019/04/05/greek-police-clash-with-refugees-heading-for-border and www.bbc.com/news/world-europe-47826607

area in Athens, evicting persons residing therein[72], prosecuting and moving some to Amygdaleza detention center. The majority of the tweets on the day refer to those events. As the situation is deteriorating in Moria camp, which hosts approximately 12,000 people, over four times its official capacity according to UNHCR, a fire breaks out on 29 September 2019, claiming the lives of two people[73]. Riots in the camp follow and tweets regarding #RefugeesGR mainly refer to the fire on the days following the event.

Following the analysis of the events around the peaks of tweet volume in 2019, the investigation turns to the hashtags of the same year. Popular location-based hashtags now include not only #Lesvos, #Moria and #Chios, as previous years, but also #Samos and #Diavata, which gained popularity by the events of April 2019 described above. The hashtag #OpenTheIslands continues to be of interest to Twitter users this year, but the hashtag #EuTurkeyDeal is no longer included among the top twenty, but rather replaced by the hashtag #StopTheToxicDeal. Moreover, solidarity-based hashtags only include #RefugeesWelcome, while a newly popular hashtag #NoBorders is making an appearance in the top twenty for 2019. Figure 4-27 below lists the twenty most popular hashtags regarding #RefugeesGR for the year 2019.

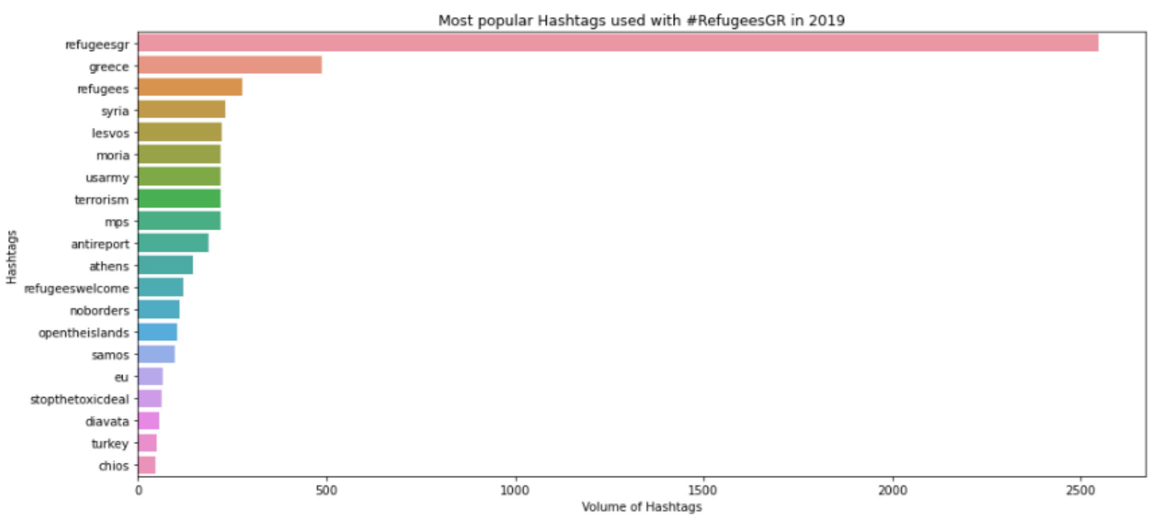

Figure 4-27: Most popular hashtags used alongside #RefugeesGR in 2019

[72] www.theguardian.com/world/2019/aug/26/greece-police-raid-athens-squats-exarcheia-arrest-migrants-agency-reports
[73] edition.cnn.com/2019/09/30/europe/lesbos-moria-refugee-camp-greece-fire-riots-intl/index.html

Similar to the years before, the words “people”, “camp”, “child” and “island” remain in the top five of words used in tweets regarding #RefugeesGR in 2019. Solidarity words continue to be less popular this year, much like 2018, with only “support” being prevalent among the most popular words. On the other side, distress words remain in the top twenty, with “strike” and “burn” being the most popular. The words “police” and “squat” also gain much more popularity this year, likely in connection to events of August 2019 described above. Figure 4-28 showcases the full picture of the hundred most commonly used words in tweets about #RefugeesGR during 2019.

Figure 4-28: Top 100 most used words in tweets with hashtag #RefugeesGR in 2019

Concluding the findings of the year, Figure 4-29 below depicts the most popular tweets on #RefugeesGR during 2019, based on the number of likes received.

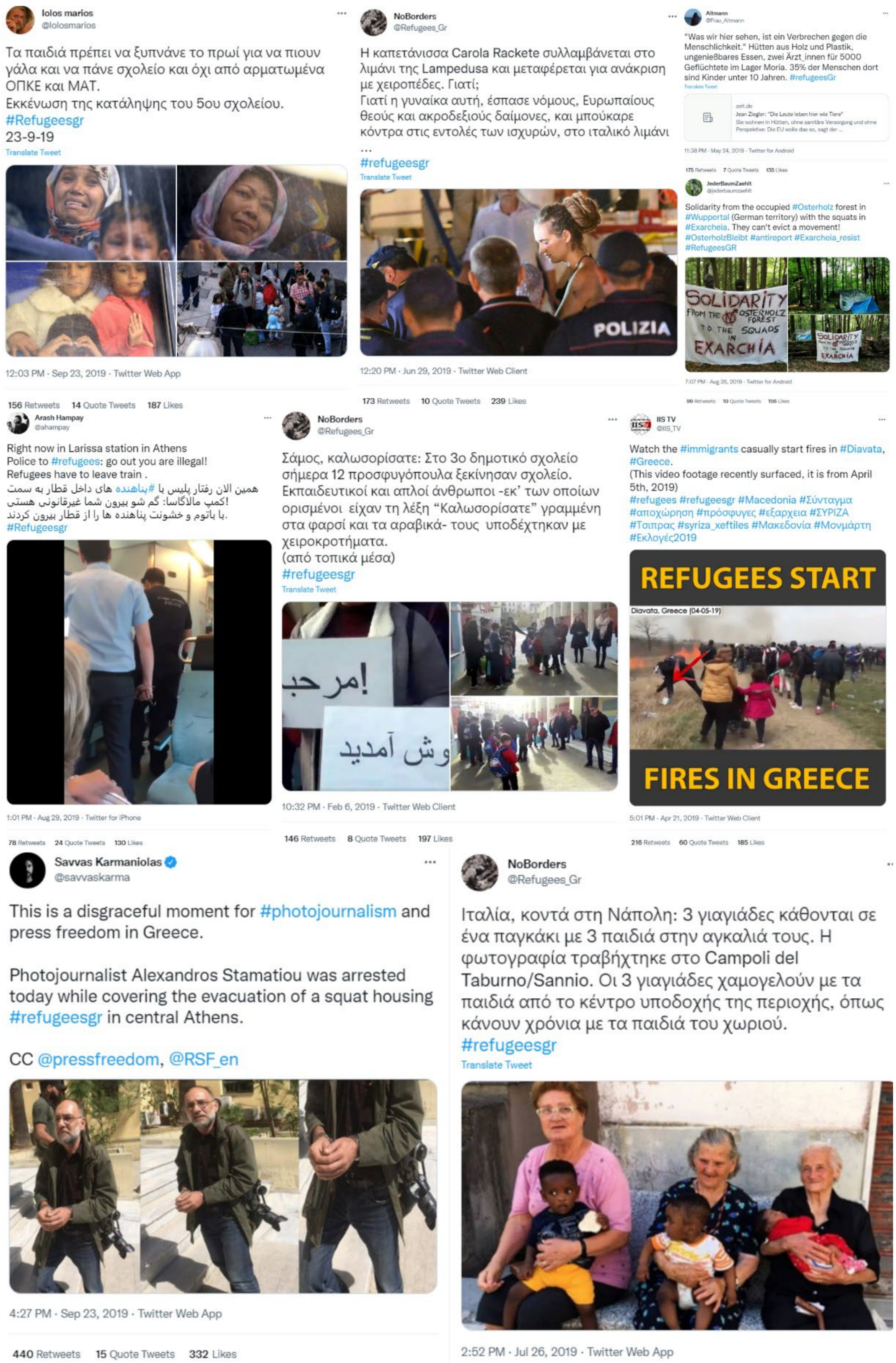

Figure 4-29: Most liked tweets in 2019

4.2.6 Year 2020

In 2020, arrivals to Greece decreased by almost 80% compared to 2019, with a little more than 15,000 persons arriving in total and about 100 dead and missing persons[74]. The low numbers of arrivals can be mainly attributed to the COVID-19 pandemic, as the arrivals trend shows significant numbers per month (more than 2,500 people) for only the first two to three months of the year. However, the proportion of persons dead and missing at sea in 2020 was significantly higher than all the previous years, increasing by 500% compared to 2019 and by 100% compared to 2018, which was the worst year in this regard up to that point.

Despite the low numbers of arrivals, the number of tweets about #RefugeesGR in 2020 increased by almost 150% compared to 2019. Three distinct peaks can be observed in Figure 4-30 among the daily volume of tweets, which will be discussed further.

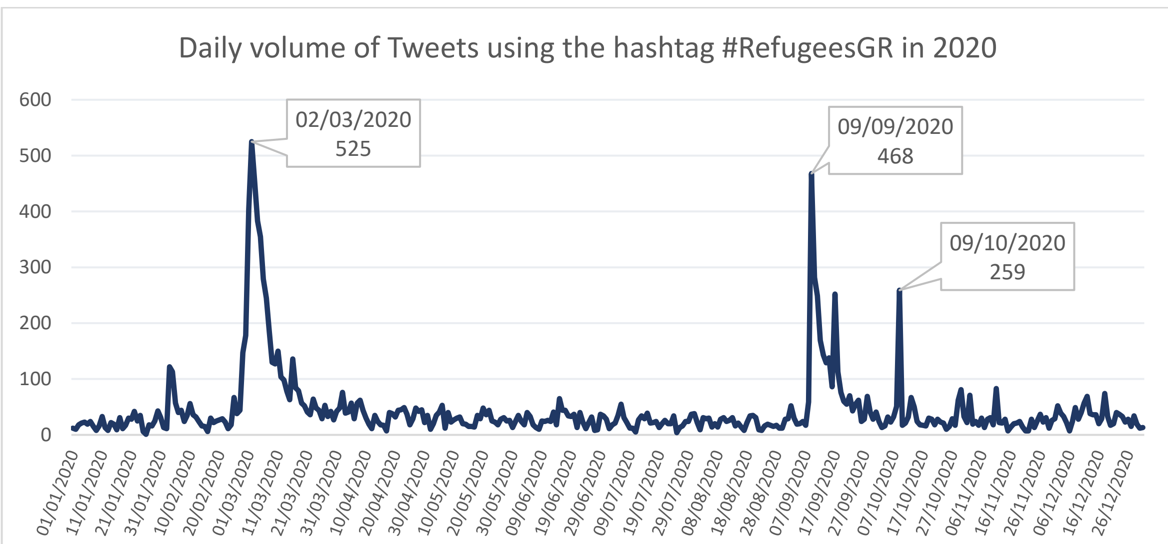

Figure 4-30: Daily volume of Tweets and peak dates in 2020

There was very high tweet activity in early March 2020 and the reasons behind the prolonged peak were not singular, as multiple events unfolded simultaneously during that time. On 28 February, Turkish government officials were quoted by media saying, “We have decided,

[74] data.unhcr.org/en/situations/mediterranean/location/5179

effectively immediately, not to stop Syrian refugees from reaching Europe by land or sea", in response to the deaths of 33 Turkish soldiers killed in an air strike in Idlib[75]. As a result, more than 10,000[76] asylum-seekers and refugees started arriving at the land border between Turkey and Greece, in Evros, and many were allowed to enter the buffer zone. Clashes with border police followed for many days[77], with rumors and inflammatory statements[78] issued on either side. Greece also introduced an emergency legislative decree[79], suspending the right to seek asylum for individuals entering Greece for the period of a month[80], a measure condemned both by UNHCR[81] and the EU commissioner for home affairs, Ylva Johansson[82], as having no legal basis. The events of the situation at the border in early March 2020 were heavily tweeted. At the same time, tensions on the island of Lesvos have been raising, with journalists, photographers, and humanitarian workers (including the head of the UNHCR Office in Lesvos) being systematically harassed or physically attacked by extremists on a daily basis. The US embassy in Greece releases an unprecedented travel advisory to US citizens on Lesvos due to the situation, while neo-Nazi groups (affiliated with the Identitarian movement) and far-right vloggers arrive on the island to capitalize on the situation[83]. Protests, clashes with police, locals preventing boats from disembarking and fires destroying NGO and civil society facilities[84] are among the most talked about topics on Twitter during that period regarding the tensions in Lesvos. On 9 September 2020, almost a year after the first large fire in Moria, a second large fire completely destroys the camp that is hosting about 13,000 asylum-seekers and refugees[85]. A month later, the new camp that was hastily erected

[75] www.aljazeera.com/news/2020/2/28/33-turkish-soldiers-killed-in-syrian-air-raid-in-idlib
[76] media sources refer anywhere between 12,000 – 30,000 people
[77] edition.cnn.com/2020/03/01/europe/turkey-greece-migrants-open-border-intl/index.html
[78] https://www.france24.com/en/20200303-erdogan-warns-europe-to-expect-millions-of-migrants-after-turkey-opens-borders
[79] www.immigration.gr/2020/03/pnp-anastolh-ths-ypovolis-aithseon-asylou.html?m=1
[80] www.nytimes.com/2020/03/01/world/europe/greece-migrants-border-turkey.html
[81] www.unhcr.org/news/press/2020/3/5e5d08ad4/unhcr-statement-situation-turkey-eu-border.html
[82] www.theguardian.com/world/2020/mar/12/greece-warned-by-eu-it-must-uphold-the-right-to-asylum
[83] www.aljazeera.com/features/2020/5/6/how-the-greek-island-lesbos-became-a-stage-for-europes-far-right
[84] www.cnbc.com/2020/03/01/refugee-crisis-in-greece-tensions-soar-between-migrants-and-locals.html
[85] www.bbc.com/news/world-europe-54082201

to replace Moria in Lesvos, Kara Tepe, floods due to rain[86], destroying almost 10% of the accommodation facilities and most tweets refer to the new camp as "Moria2".

After looking into the event analysis for 2020, the discussion turns to hashtags, in order to validate the consistency of the topics. Perhaps unsurprisingly, the most popular location-based hashtags this year include #Lesvos and #Moria, similar to 2019, but also #Evros, which is on par with the events of March 2020, as described above. Solidarity-based hashtags this year also include #RefugeesWelcome, much like the year before, but the hashtag #LeaveNoOneBehind is newly trending in 2020. Furthermore, two new hashtags enter the top twenty most used in relation to #RefugeesGR this year: one is #Covid19, relating to the pandemic, and the other is #Pushbacks. The full list of the most popular hashtags for 2020 is summarized in Figure 4-31.

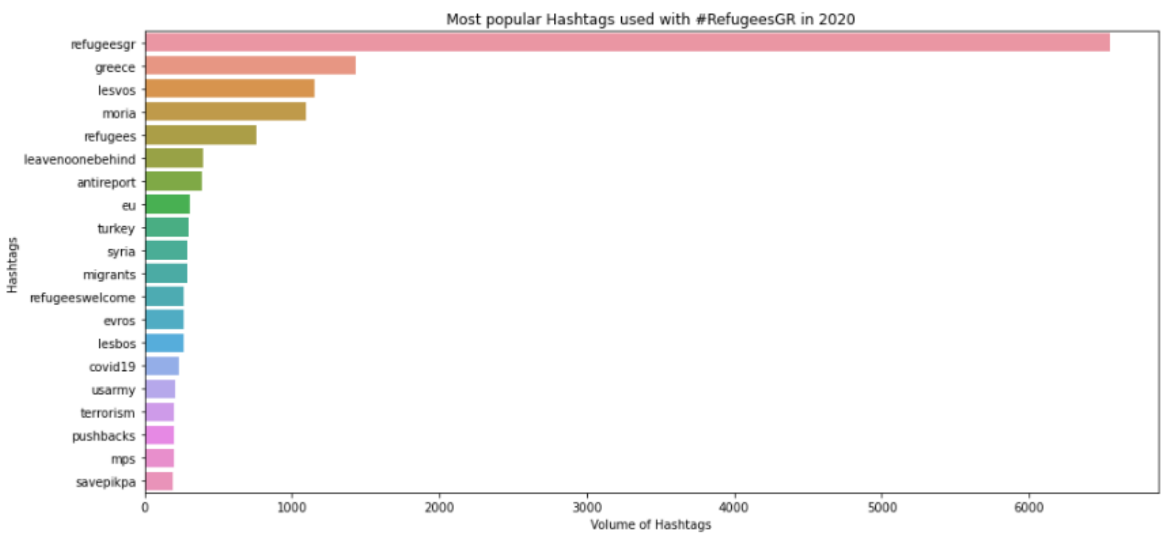

Figure 4-31: Most popular hashtags used alongside #RefugeesGR in 2020

Much like 2019, the words "camp", "people" and "island" continue to show in the top five most used words about #RefugeesGR in 2020. However, the word "asylum" has climbed to the top this year as well. The word "border" also makes a comeback in the top twenty words, relating to the events of March 2020, and for the first time the word "government" is also widely used in tweets. Words that relate to the fire in Moria, "fire" and "burn", are also quite

[86] www.dw.com/en/lesbos-is-another-moria-in-the-making/a-55249863

popular in 2020, as is “police” and “migrant”, similar to the previous years. The word “right” also appears in the top twenty most used words, which many tweets used as part of the bigrams “far right” or “right wing”, but also on its own to note “access to rights”. The full one hundred most used words in tweets about #RefugeesGR in 2020 are shown in Figure 4-32 below.

Figure 4-32: Top 100 most used words in tweets with hashtag #RefugeesGR in 2020

Wrapping up the yearly summary, Figure 4-33 below depicts the most popular tweets on #RefugeesGR during 2020, according to the number of likes received.

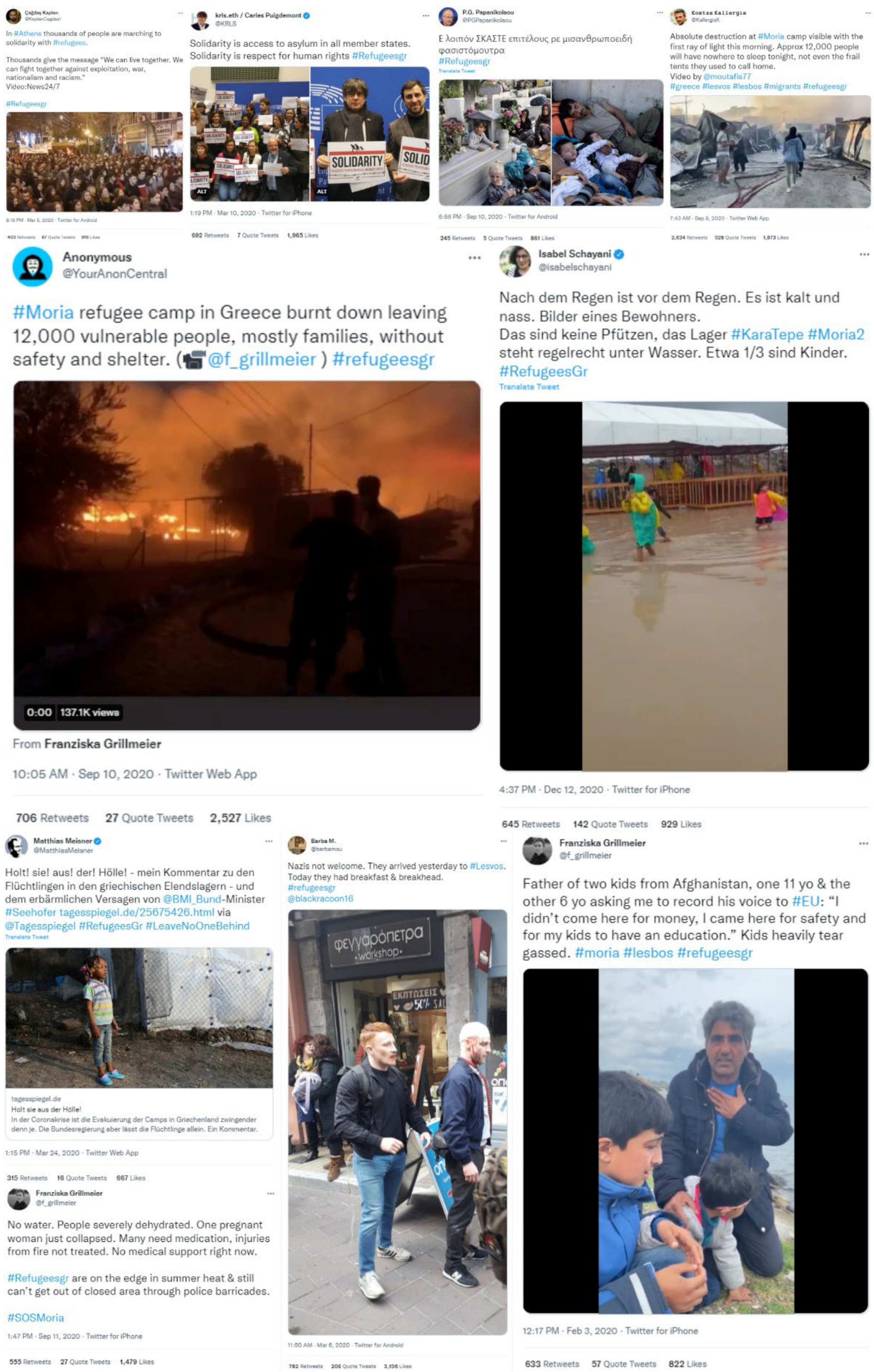

Figure 4-33: Most liked tweets in 2020

4.2.7 Year 2021

New arrivals in Greece during 2021 further decreased by 40% compared to 2020, with only about 9,000 persons arriving in total and 115 dead and missing persons[87]. Despite the number of arrivals being quite minimal compared to previous years and well below 1,000 persons per month for much of the year, the proportion of persons dead and missing at sea skyrocketed to 1.3% in 2021, increasing by more than 110% compared to 2020.

The number of daily tweets about #RefugeesGR in 2021 is on par with the low numbers of arrivals, with a 60% decrease compared to 2020. There is only one significant peak that stands out in 2021, as Figure 4-34 suggests, however a few more peaks that are close to 50 tweets per day are marked for discussion purposes.

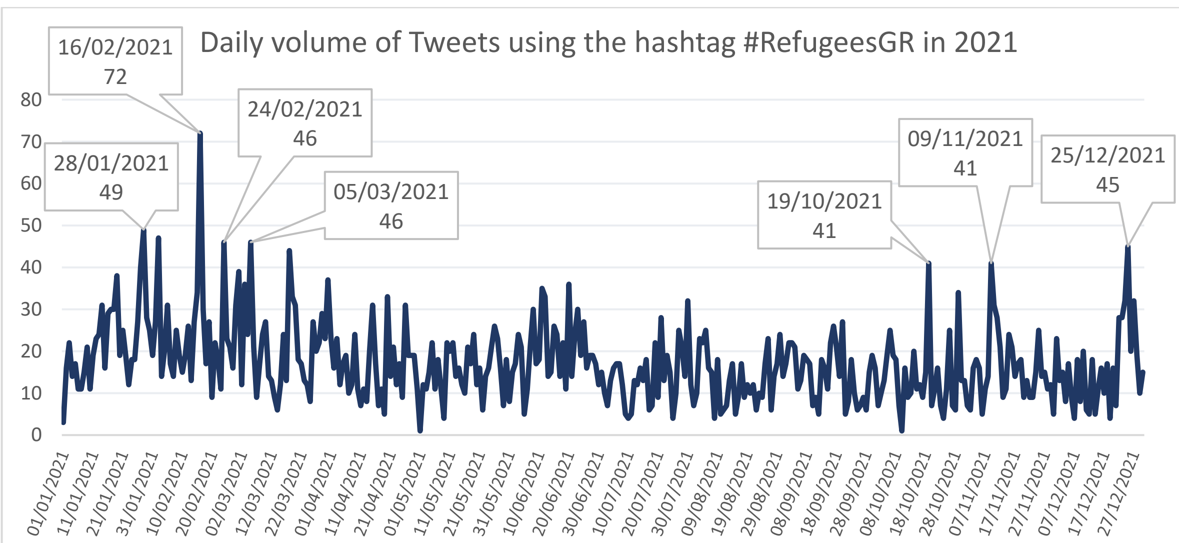

Figure 4-34: Daily volume of Tweets and peak dates in 2021

Since tweet volume is very low during 2021, it’s difficult to separate the most tweeted events of that year simply by numbers. On late January 2021, tweets mainly discuss reports confirming lead contamination of the soil in the new camp in Lesvos, Kara Tepe[88]. On early

[87] data.unhcr.org/en/situations/mediterranean/location/5179
[88] http://www.hrw.org/news/2021/01/27/greece-migrant-camp-lead-contamination

February, as the weather becomes colder, tweets draw attention to harsh conditions in the camps on Lesvos, Chios and Samos[89]. When extreme weather conditions finally hit Greece in mid-February, Malakasa, Schisto and Eleonas camps were completely covered in snow[90] with no access to heating or electricity and residents had to be eventually evacuated. Tweet volume also peaked on 24 February, after reports on the death of a seven-year-old boy in a fire at a camp in Thebes[91] and the attempted suicide of a pregnant woman in Kara Tepe, in Lesvos, who was charged with arson[92]. Tweets on early March mainly discuss the risk of homelessness for recognized refugees in Greece[93], as the Filoxenia program is being phased out without an alternative in its place. At the same time, focus is shifted towards the investigation that the EU Border Agency, FRONTEX, faces under the newly established "Frontex Scrutiny Working Group" of the European Parliament[94], which is tasked to investigate pushback allegations in Greece and FRONTEX's involvement in those[95]. On 19 October, RSA publishes a timeline report on the removal of a group, including a Syrian asylum-seeker at risk of refoulement, from Greece and their return to Turkey[96] by Greek authorities. On 9 November, most tweets discuss the verbal altercation between Greek Prime Minister, Kyriakos Mitsotakis and Dutch journalist, Ingeborg Beugel, when the reporter inquired about pushback allegations in Greek territory[97]. The final peak for the year 2021 in tweet volume was on 25 December and concerned a shipwreck off of Antikythera island that left nine dead[98].

---

[89] medium.com/are-you-syrious/another-surprise-winter-on-the-islands-chios-samos-and-lesvos-36dabd5285ec

[90] asylumineurope.org/reports/country/greece/reception-conditions/housing/conditions-reception-facilities

[91] www.amna.gr/en/article/531488/Blaze-at-migrant-camp-in-Thebes-seven-year-old-boy-dead

[92] www.theguardian.com/world/2021/feb/26/woman-who-set-herself-on-fire-in-lesbos-refugee-camp-may-face-arson-charges

[93] www.theguardian.com/world/2021/mar/05/greece-thousands-of-migrants-at-risk-of-homelessness-as-eu-scheme-ends

[94] www.euronews.com/my-europe/2021/02/24/meps-to-personally-investigate-frontex-amid-pushback-allegations

[95] Eventually the Executive Director of FRONTEX, Fabrice Leggeri. resigns in April 2022, following more reports in the involvement of FRONTEX in pushbacks in Greece. www.lemonde.fr/en/international/article/2022/05/01/the-story-behind-frontex-director-fabrice-leggeri-s-resignation_5982123_4.html

[96] rsaegean.org/en/timeline-pushback-evros

[97] www.theguardian.com/world/2021/nov/10/greek-prime-minister-angrily-defends-treatment-of-refugees

[98] www.aljazeera.com/news/2021/12/24/seven-people-killed-in-shipwreck-off-greek-islet

Having concluded the exploration of events that sparked Twitter's interest in 2021, the next step is to understand the more popular hashtags accompanying these tweets. Location-based hashtags are not included amongst the top five most popular ones, like it was the case in all previous years, but rather #Lesvos, #Samos, #Moria2 and #Moria appear further down in the list this year. Much like 2020, solidarity-based hashtags this year include #RefugeesWelcome and #LeaveNoOneBehind, although in lower frequencies, as also the overall tweet volume is lower this year. The hashtag #Pushbacks, that made its first appearance in the top twenty most popular hashtags regarding #RefugeesGR in 2020 is now in the third place of popularity for 2021. Figure 4-35 below portrays the full list of the most popular hashtags on #RefugeesGR for 2021.

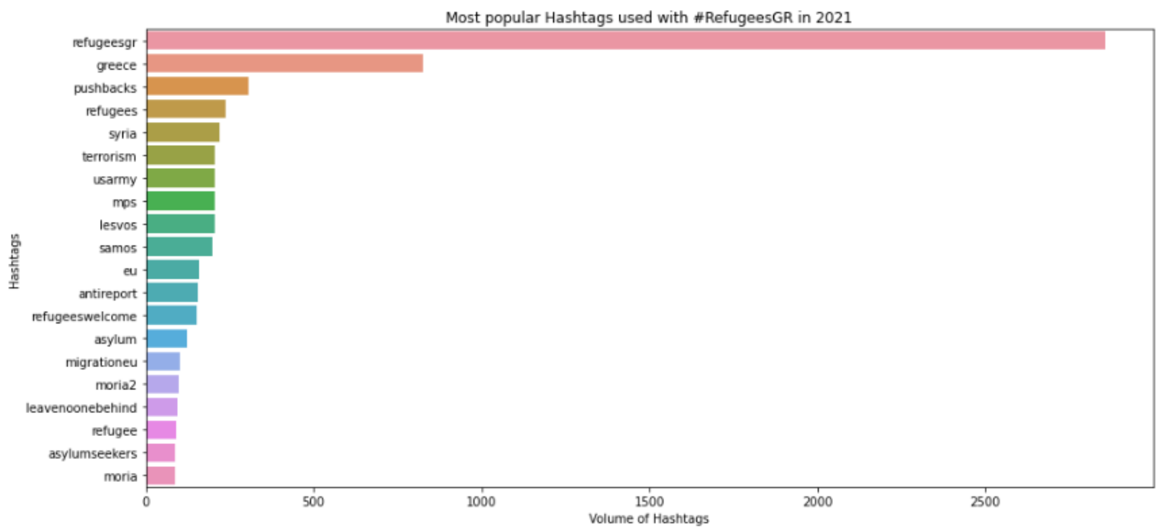

Figure 4-35: Most popular hashtags used alongside #RefugeesGR in 2021

Similar to the four previous years, the words "camp" and "people" are among the top two most used words regarding #RefugeesGR in 2021. Also similar to 2020, the word "asylum" is among the top five, but in 2021 the word "seeker" has also climbed to the top twenty. The word "border" remains popular, like it also was in 2020, as is the word "right" that first appeared so high in popularity in 2020. The word "police" is not as popular in 2021 as it was the previous year, but the word "authority" has gained popularity instead. Figure 4-36 below portrays the full picture of the most commonly used words in tweets about #RefugeesGR in 2021.

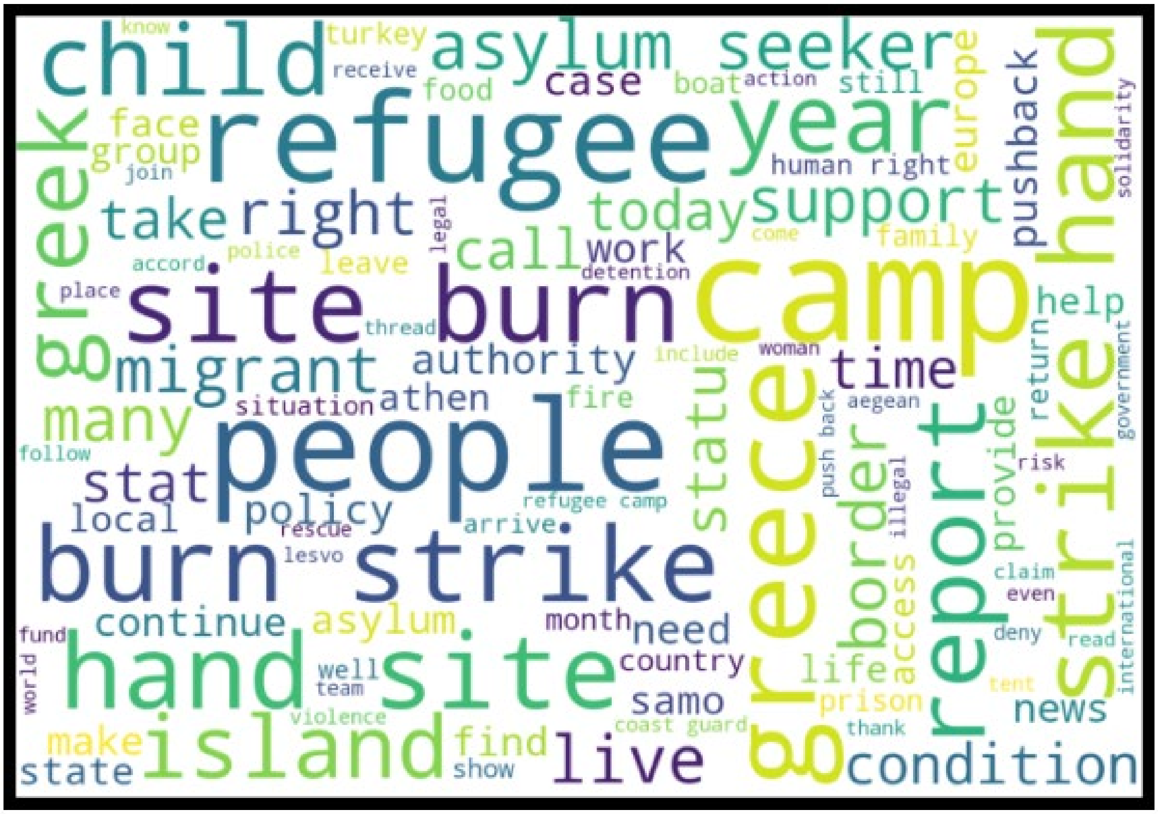

Figure 4-36: Top 100 most used words in tweets with hashtag #RefugeesGR in 2021

Finalizing the yearly findings, Figure 4-37 below provides a collage of the most popular tweets on #RefugeesGR during 2021, based on the number of likes the tweets received.

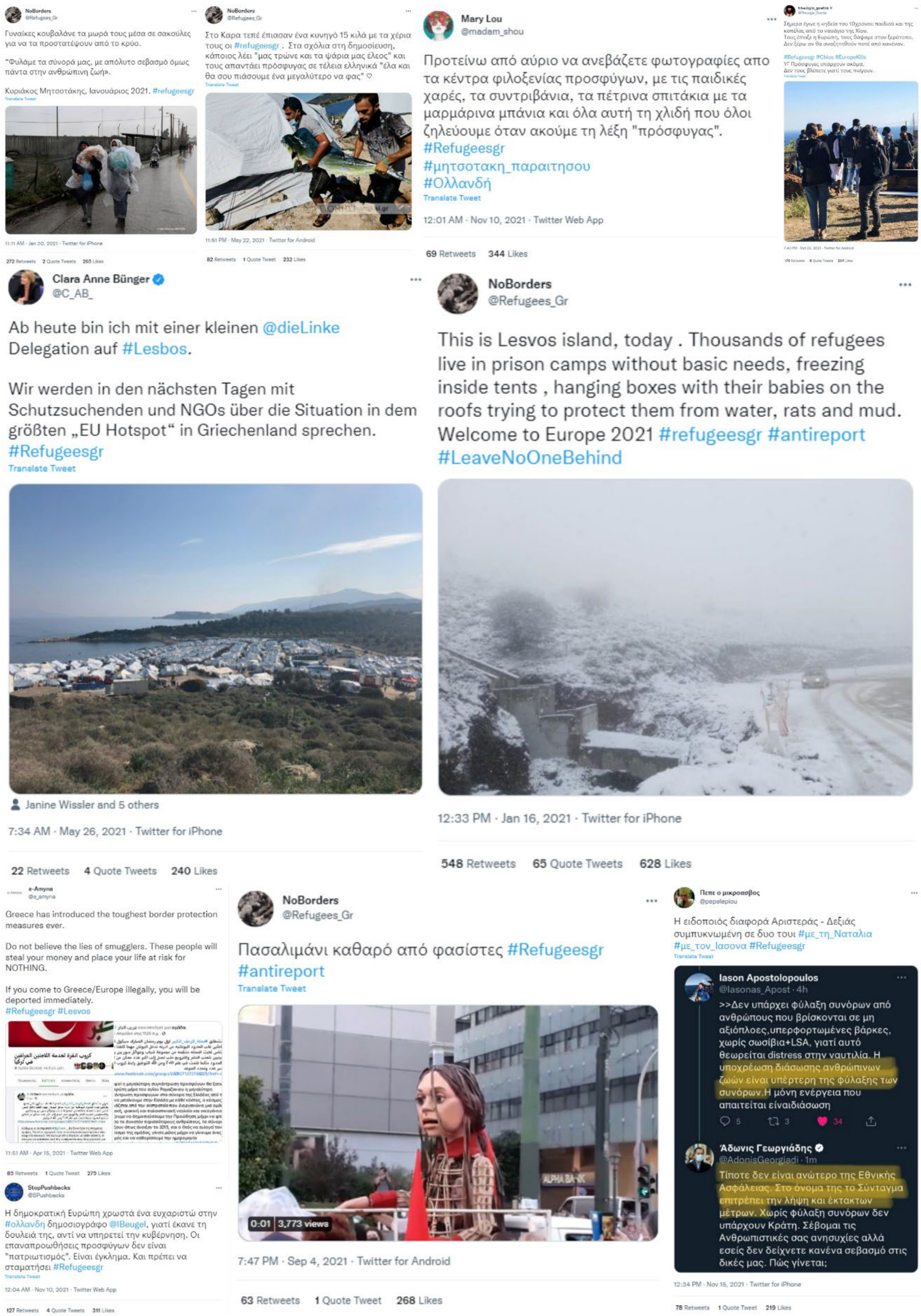

Figure 4-37: Most liked tweets in 2021

4.3 Cleaning tweets

In Chapter 3.3, the methodology of cleaning and pre-processing tweets was described in detail, utilizing the python library “re” for preprocessing and reducing noise, “nltk” for removing stopwords and “spacy” for lemmatization.

In this chapter, the results of the described actions will be showcased with a real tweet example[99] by the account of the EU Commission for Civil Protection and Humanitarian Aid (@eu_echo) from 12 August 2017, as shown in the Figure 4-38 below.

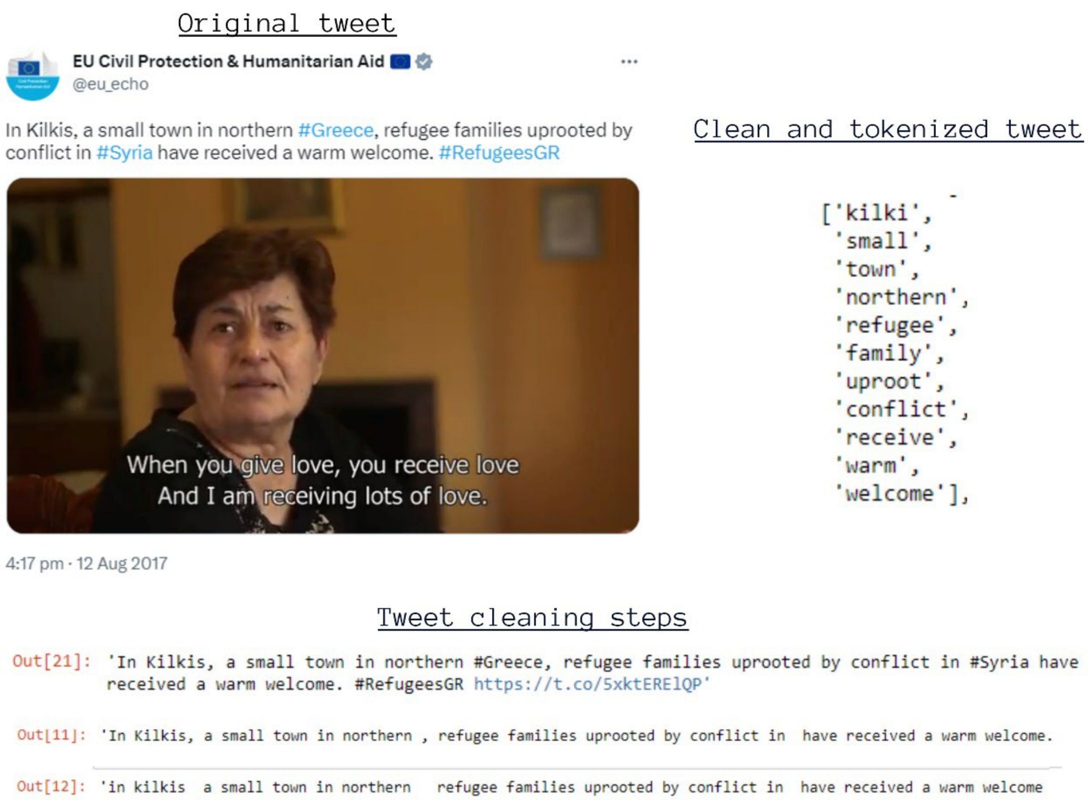

Figure 4-38: Tweet cleaning example[99] using a real tweet from the dataset, posted by @eu_echo

On the top left of Figure 4-38 above, a snapshot of the original tweet is shown, while the text of the tweet, as mined into the dataset is shown in output 21. The original tweet reads: “In Kilkis, a small town in northern #Greece, refugee families uprooted by conflict in #Syria

[99] Permalink for the tweet can be found here: https://twitter.com/eu_echo/status/896359878876307458

have received a warm welcome. #RefugeesGR”. The tweet also includes a video from the ESTIA accommodation programme in Kilkis, which was implemented at the time by the local NGO “Omnes” and was funded by UNHCR and EU ECHO. The first preprocessing step for the tweet is the removal of hyperlinks, mentions and hashtags and the result of this step is shown in output 11. Specifically, the video link is removed, along with hashtags #Greece, #Syria and #RefugeesGR. The second part of tweet preprocessing includes conversion of characters to lowercase, and removal of all punctuation and numbers, while the result is shown in output 12, reading “in Kilkis a small town in northern refugee families uprooted by conflict in have received a warm welcome”. The last step of the cleaning process is the removal of stopwords from the tweet, which is then lemmatized and tokenized, with the result being visible on the top right of Figure 4-38. In this tweet example, the stopwords “in”, “a”, “by” and “have” were removed altogether. The verbs “uprooted” and “received” were normalized through the lemmatization process and were transformed to their root forms, namely “uproot” and “receive”. Moreover, the word “families” was transformed to its singular form “family”, while the word “kilkis”, which refers to a town in northern Greece, was not recognized as such and was transformed to “kilki”. Overall, the end result of the cleaning process for this tweet does maintain the essence of its message, along with the main sentiment.

4.4 Topic Analysis

Moving on to topic modelling, this chapter will focus on investigating the output of the LDA algorithm per year for the period 2015 – 2021. For each year, a topic discussion will follow, which will highlight the most important findings of the model, based on the output of word probabilities and the visual representation on LDAvis.

Before diving into the analysis of topics for the year 2015, it is useful to showcase how the number of topics per year was selected for each LDA model. Figure 4-39 below provides a visual representation of how the chosen coherence measure, c_v, fluctuates based on the number of topics that are fed into the LDA model, as described in Chapter 3.5.2. Based on the coherence scores in Figure 4-39, 13 topics were selected as the optimal number for the 2015 LDA model. A similar distribution of coherence scores was calculated for all yearly datasets and the number of topics was selected based on the optimal score on one hand, but also human interpretation of the topics formed.

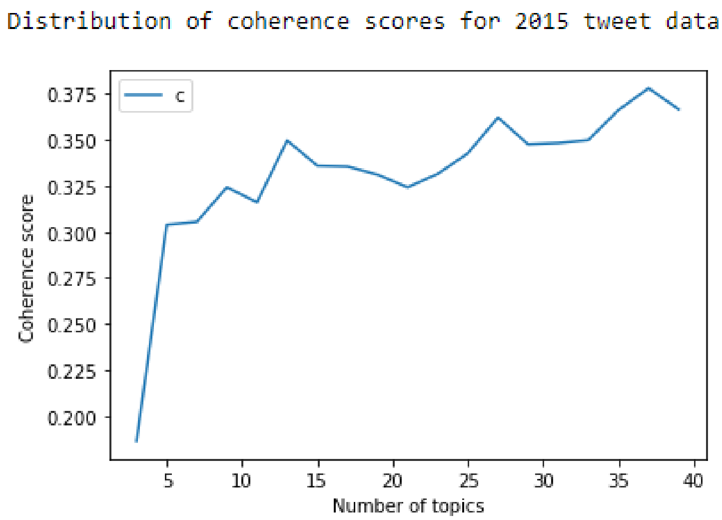

Figure 4-39: Coherence scores (c_v) of LDA for 2015 tweet data, based on number of topics

Moving on to the analysis of the topics discovered for 2015 tweet data, there are 13 topics visualized in Figure 4-40, three of which are larger than the rest and more distinctly separated, namely topics 1, 2 and 3 (numbers of topics correspond to the numbers inside the bubbles of said topics). Starting with the first quadrant, topic 2 encapsulates the notion of first arrivals of refugees in Greece, being populated with words such as: “Greek”, “island”, “arrive”, “boat”, “dinghy”, “rescue”, “photo” and “migrant”. Within the same quadrant, topic 10 refers to border crossings towards North Macedonia, with its main words being “border”, “cross”, “Greek” and “FYROM”. Topic 1 seems to be a more route-based topic about the journey of refugees in Greece, mainly populated by the word “refugee”, followed by “Greece”, “Athens”, “Syrian”, “Lesvos” and “Thessaloniki”. This topic seems to track the route that refugees, in their vast majority Syrians during 2015, were following in Greece, from the first point of arrival until their exit from the northern borders of Greece. Advancing to the third quadrant, topic 3 represents solidarity, with its main words being “help”, “volunteer”, “people”, “need”, “solidarity” and “support”. Within the same quadrant, albeit much smaller, topic 12 also seems to be about support and action, although through a different angle, with its main words being “crisis”, “Europe”, “urgent”, “send”, “asylum”, “need” and “protect”.

The rest of the topics are smaller and less distinct from one another, concentrating within the fourth quadrant. Topics 4, 6 and 8 are somewhat overlapping and seem to be all referring to life within camps and informal campsites. Specifically, topic 8 encompasses the situation within camps with prominent words being: “camp”, “many”, “water”, “shelter” and even “moria”, while topic 4 is about police interventions, with prominent words being: “police”, “report”, “stop”, “tent” and “sleep” and topic 6 focuses specifically on women and children, who are often referred to separately as they usually have specialized needs. The remaining topics cannot be as easily categorized, but topic 5 refers broadly to world news, topic 7 specifically mentions Afghan refugees, topic 11 encapsulates arrival stories and topic 13 status updates. Topic 2 is highlighted in Figure 4-40 below, so that the most relevant terms for this topic can also be reviewed in detail.

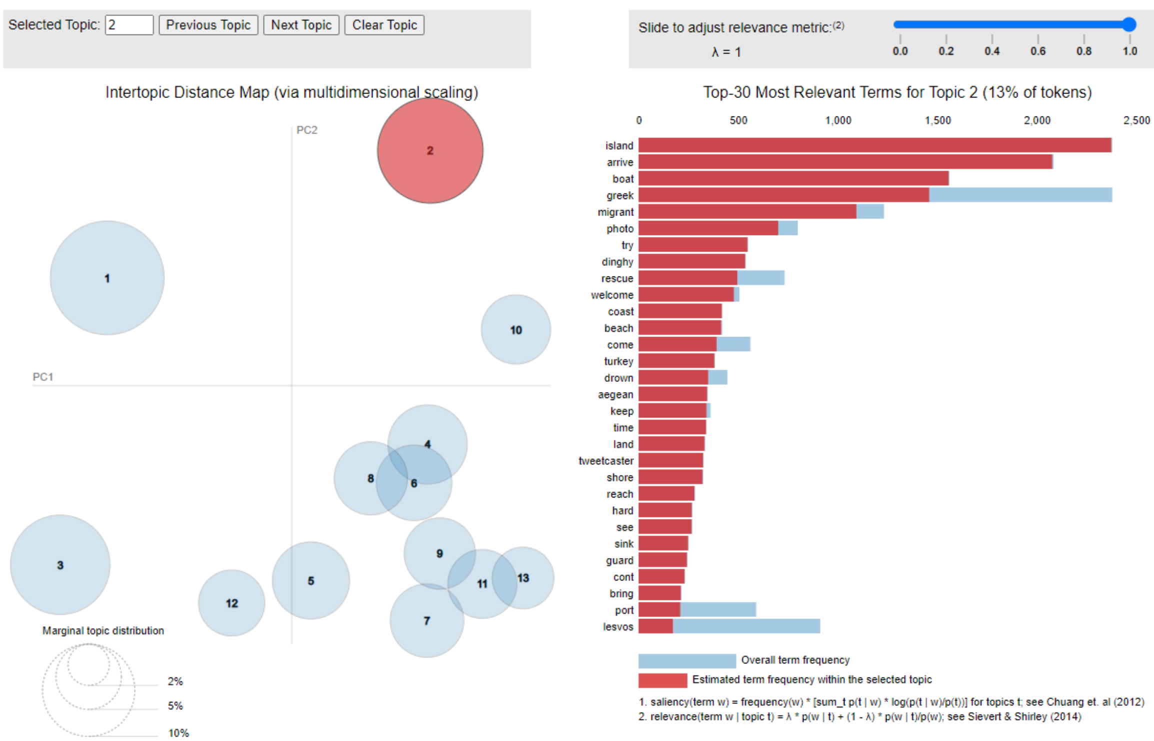

Figure 4-40: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2015

Twitter discussion about #RefugeesGR in 2016 is encapsulated within 9 topics, which are visualized in Figure 4-41. Six of those topics are concentrated closely together in the first quadrant and there seems to be significant correlation between them. Starting with topic 1 in the second quadrant, which is also the largest topic, it covers the situation in Idomeni camp at the border of Greece with North Macedonia, with some of its main words being: “refugee”, “camp”, “border”, “police”, but also location-based words such as “Idomeni”, “FYROM” and “Macedonian”. Topic 2 seems to mainly revolve around solidarity, with some main words being: “people”, “Europe”, “port”, “volunteer”, “support”, “food”, “solidarity” and “Piraeus”, which is also in line with 2016 event analysis regarding the situation in Piraeus port, where refugees were residing for a prolonged period of time. The most relevant keywords for topic 3 seem to be about daily news, updates and reports, including words such as “photo” and “report”. Regarding the six clustered topics in the first quadrant, topic 4 primarily covers call-to-action, with prominent words being: “help”, “needed”, “island”, “arrival”, “Lesvos” and “boat”. Similarly, topic 6 also seems to be about new arrivals in Greece, with prevalent words being “morning”, “boat”, “human”, “right” and “arrive”. Topic 5 cannot be easily placed within one category, but Björn Kietzmann’s photojournalistic work in Greece is prominent within this topic. Topic 7 revolves around the relocation program which started being implemented in 2016, with words such as: “asylum”, “seeker”, “hope”, “family”, “story”, “relocation”, “Germany”. Topic 8 showcases the negative side of closed accommodation facilities, with prominent words including “video”, “attack”, “detention”, “center”, “prison”, “fascist” and “moria”. Lastly, topic 9 is also not clearly defined, but could be broadly described as information sharing, with words including “update”, “info”, “UNHCR”, “Amnesty” and some individual journalists as well. Topic 7 is highlighted in Figure 4-41 below, so that the most relevant terms for this topic can be examined closely.

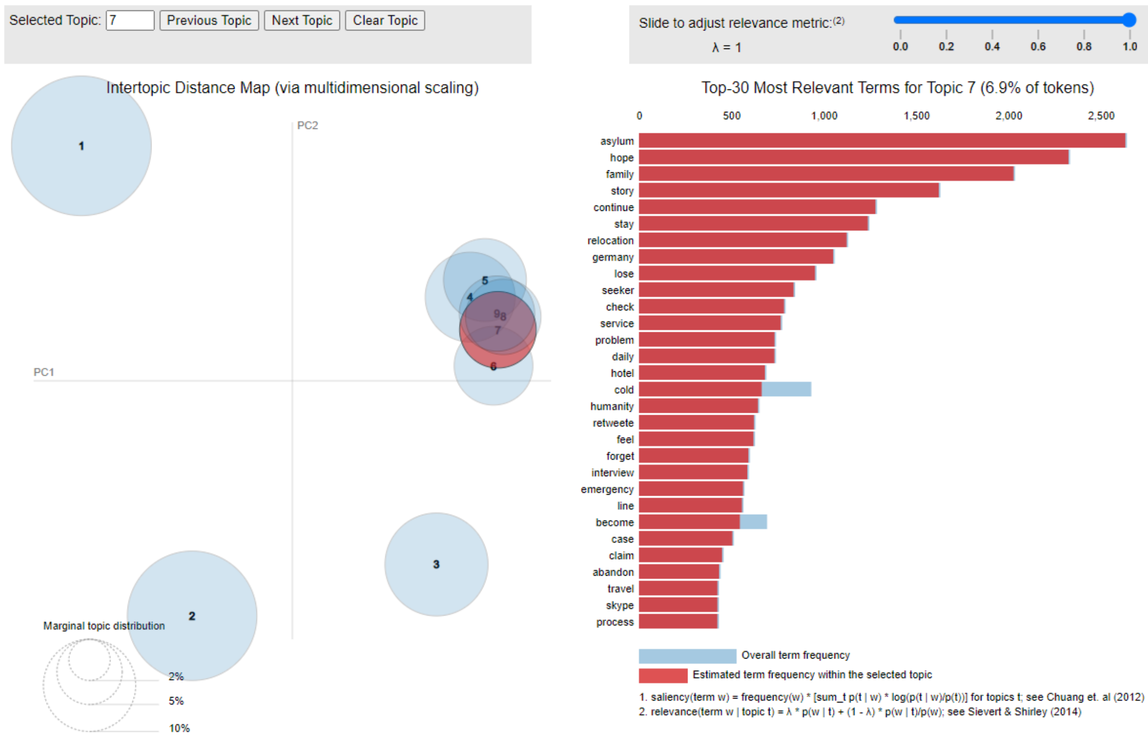

Figure 4-41: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2016

In 2017, there is much more variation present in the twelve topics discovered by the LDA model compared to 2016, although the overall number of tweets in 2017 is significantly less than 2016, as showcased in Chapter 4.2. Starting with the first quadrant in Figure 4-42, topic 11 refers to squats and evictions from them, with prominent words being: "humanity", "housing", "squat", "eviction", "solidarity" and "attack". Topics 10 and 12 are closely correlated, with topic 10 referring to detention centres and topic 12 not having a very clear single message, but a few of its words are about unaccompanied minors. Topic 13 also does not seem to have a clear overarching theme, but suicide attempts are distinctly encapsulated within that topic. Similar to previous years, there is a topic clearly referring to new arrivals on Greek islands, namely topic 3 for 2017. Topics 6, 7 and 8 are closely correlated with each other, but overall do not have a high concentration of words. Topic 6 describes the difference between persons living in the islands and the mainland, with some prominent words being "hope", "move" and "go". Topic 7 is about education for refugee children and topic 8 about Turkey and deportations there. Topic 1 refers to the conditions in refugee camps in Greece, including words such as “moria” and “winter”, which was amongst the themes driving tweet volume during 2017, as described in Chapter 4.2.3. Closely correlated with topic 1, topic 5 is about refugee rights on Greek islands, also including words about the conditions there. Both topics 2 and 9 are generally about solidarity and call-to-action and perhaps should be more closely correlated with each other. Topic 2 specifically includes words such as “hunger”, “strike”, “need”, “solidarity”, “help”, “volunteer” and “support” and topic 9, which is highlighted in Figure 4-42 below so that it can be fully reviewed, also includes “help”, “support”, but also “thank” and “Anwar Nillufary”[100]. Finally, topic 4 refers more broadly to opening the borders of Greece.

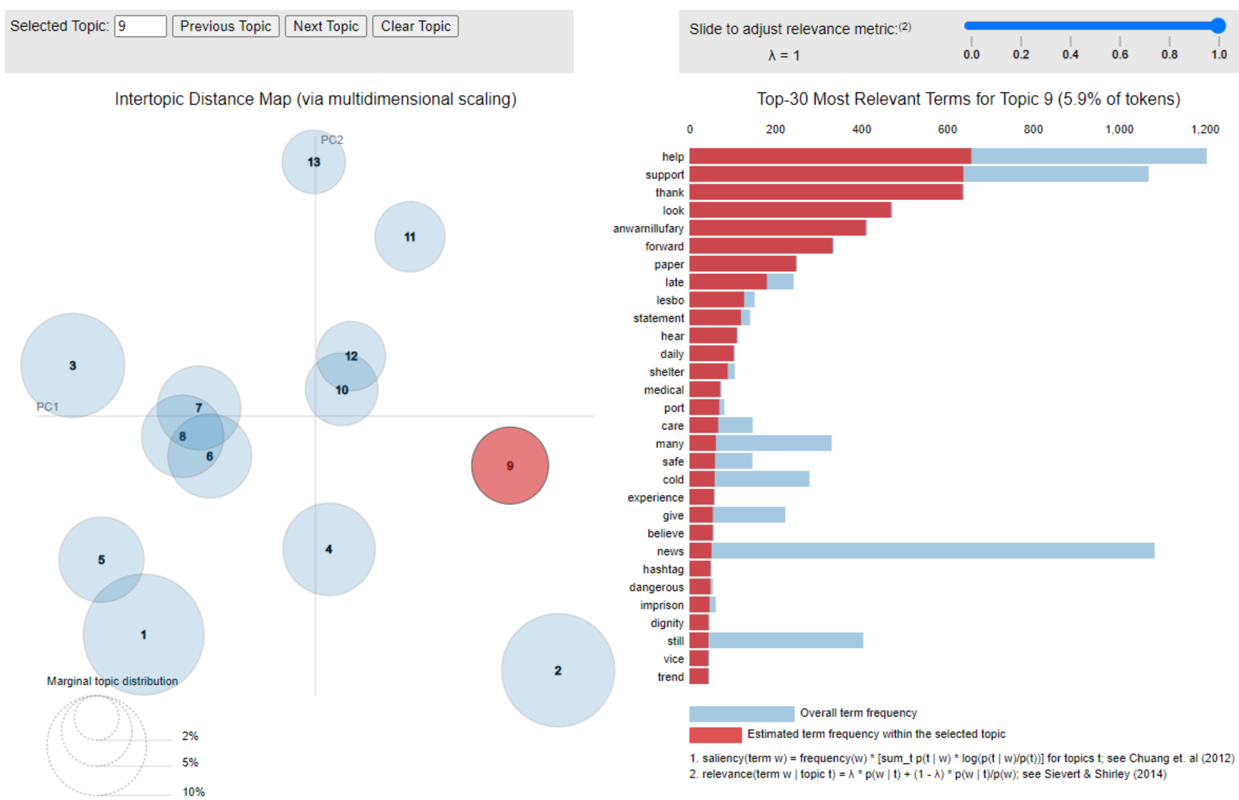

Figure 4-42: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2017

[100] Anwar Nillufary is a Kurdish refugee, who went on hunger strike outside the UNHCR Office in Athens in 2017 requesting resettlement kurdistanhumanrights.org/en/greece-kurdish-refugee-on-hunger-strike-outside-unhcr-office-demands-resettlement

As Figure 4-43 below suggests, there are nine separate topics derived from LDA for 2018, with four of them being more closely clustered together. Starting from the first quadrant, topic 4 encapsulates solidarity efforts, with most prominent words being “refugee”, “support”, “help”, “provide” and “asnteamuk”, which refers to the Aegean Solidarity Network[101], a UK charity active in Greece. Both topics 6 and 7 refer to conditions in refugee camps in Greece, and specifically Moria camp, but topic 6 emphasizes more the story of residents and topic 7 the safety aspect. Looking into the cluster of topics in quadrant 2, topic 8 refers to refugee protests and attacks, while topic 3, which is much larger, encapsulates more broadly the attacks that took place in Lesvos island during 2018 by far-right extremists, as also described in the event analysis of Chapter 4.2.4. Topic 1, which is the one highlighted in Figure 4-43 and can be explored further, refers to camp conditions and specifically includes the word “Moria”, but also encapsulates the violent events and general unrest in the camp during 2018 on the island of Lesvos. Topic 2 depicts boat arrivals under a single theme, similar to previous years, with its main words being “people”, “boat”, “woman”, “arrive”, “rubber”, “shore” and “land”. Finally, topic 5, although not very clearly monothematic, includes several words referring to the “mental” and “psychological” situation in refugee camps.

[101] www.youtube.com/watch?v=MCiH4UdpBz0

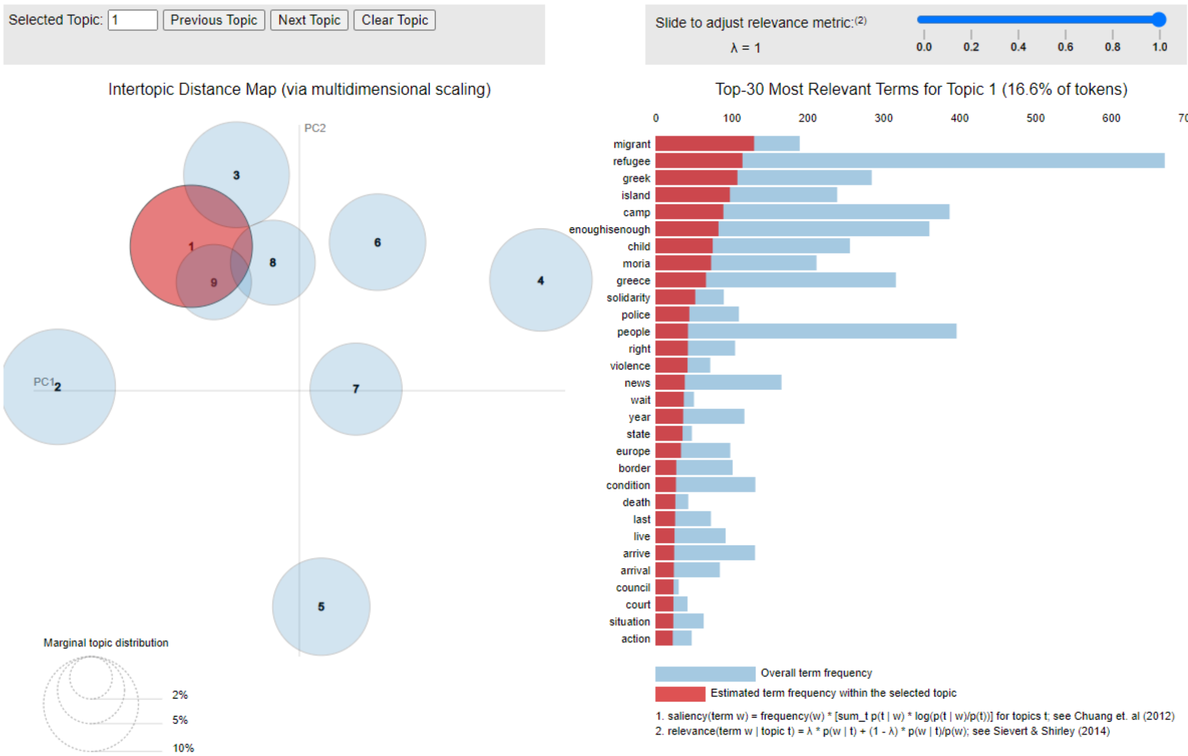

Figure 4-43: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2018

Similar to the previous year, tweet volume during 2019 regarding #RefugeesGR was deflated, therefore the volume of words per topic is also quite low and correlations might not be as distinct. Starting from topic 10 in the first quadrant, it only significantly includes the phrase "enoughisenough", a popular hashtag used to express frustration, although it also incorporates words referring to "refugee" and "squat". Similarly, topic 8 only significantly includes the word "police", in relation to "camp", "protest" and "border". Moving into the cluster of topics in the second quadrant, topic 2 refers to refugee evictions, mainly regarding "squats", but also including words such as "camps", "housing" and "homeless". Topic 2 is also highlighted in Figure 4-44 below, so that the most relevant terms for this topic can be reviewed in detail. Topic 3 seems to be about the 2019 fire in Moria camp, while topic 4 revolves around refugee arrivals. Topic 1 mostly encapsulates the conditions in camps on the Aegean islands, including specific words such as: "moria" and "capacity". Topics 5 and 6 overlap quite a bit, with topic 5 referring to the rights of asylum seekers and refugees in Greece, while topic 6 focuses on support and help for refugees, including words like "human" and “right”. Similar to topic 4, topic 7 also mentions boat arrivals but more in relation to “Turkey” and “rescue” efforts. Lastly, topics 11 and 9 do not show clear indications of a single theme, based on their most popular terms.

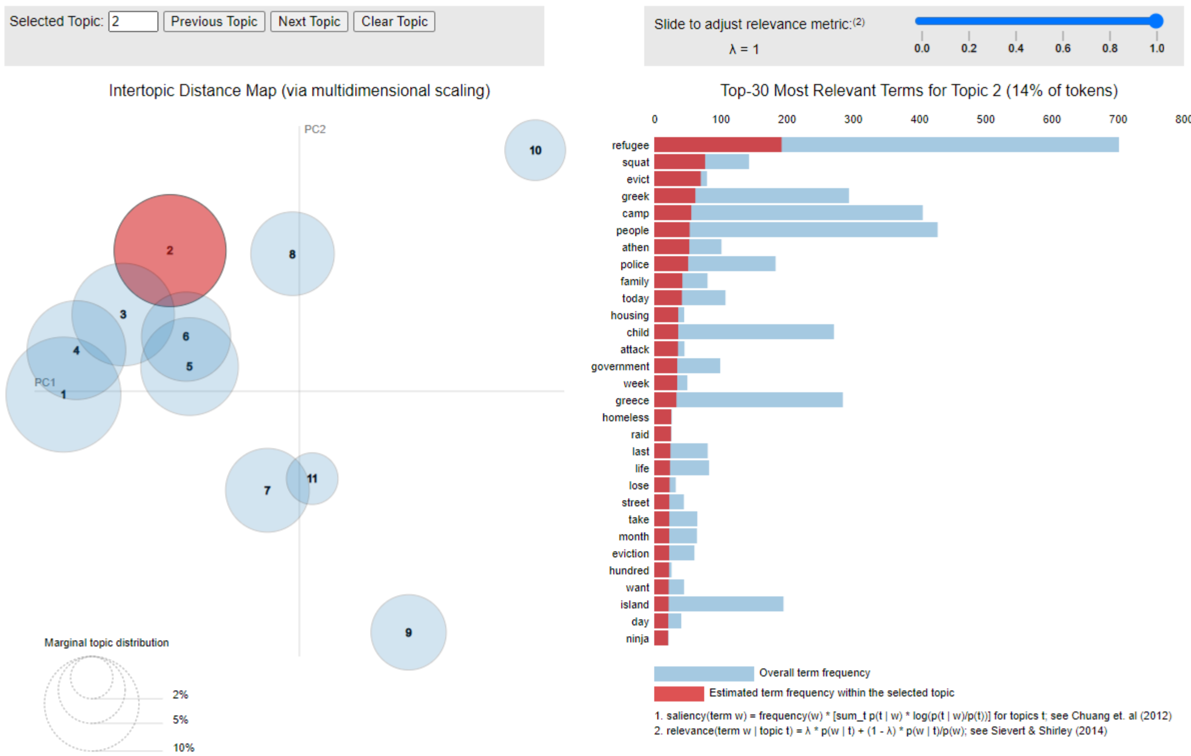

Figure 4-44: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2019

For 2020, there is a total of 11 topics derived from the LDA model and the majority of them show concentration on the second and third quadrants. Starting from topic 1, which is also the one highlighted in Figure 4-45 below, the topic is generally referring to conditions in Moria camp in Lesvos, but also specifically includes the word “fire”, as 2020 was the year that Moria camp was destroyed by fire, an event included in the analysis for that year in Chapter 4.2.6. Topic 2 is mostly relating to detention centres, with some of the most prominent words being “camp”, “island”, “detention”, “centre”, “close”, “detain”, and “prison”. Topic 6 refers to boats at the Greek sea borders, with some of the most relevant words including “coast”, “guard”, “rescue”, “Turkish” and “pushback”. Topic 8 is also relating to the Greek islands, but specifically on the events of March 2020 on Lesvos and the attacks by far-right extremists, which are described in the event analysis in Chapter 4.2.6.

Topics 9 and 10 both refer to the Greek borders, specifically encapsulating the events of March 2020 on the borders between Greece and Turkey in Evros, as the main words therein are “police”, “force”, “tear” and “Turkey”. In addition, topic 7 also refers to the Greek-Turkey border situation, albeit in less aggressive terms, using words such as “crisis”, “human”, “right”, refugee”, “flee”, “solidarity” and “humanitarian”. Topic 3 seems to have a collection of call-to-action words regarding rights of asylum seekers and refugees and topic 5 solely relates to the PIKPA[102] facility in Lesvos, which used to host very vulnerable asylum-seekers and refugees. Finally, the most used terms within Topics 4 and 11 do not represent a single theme.

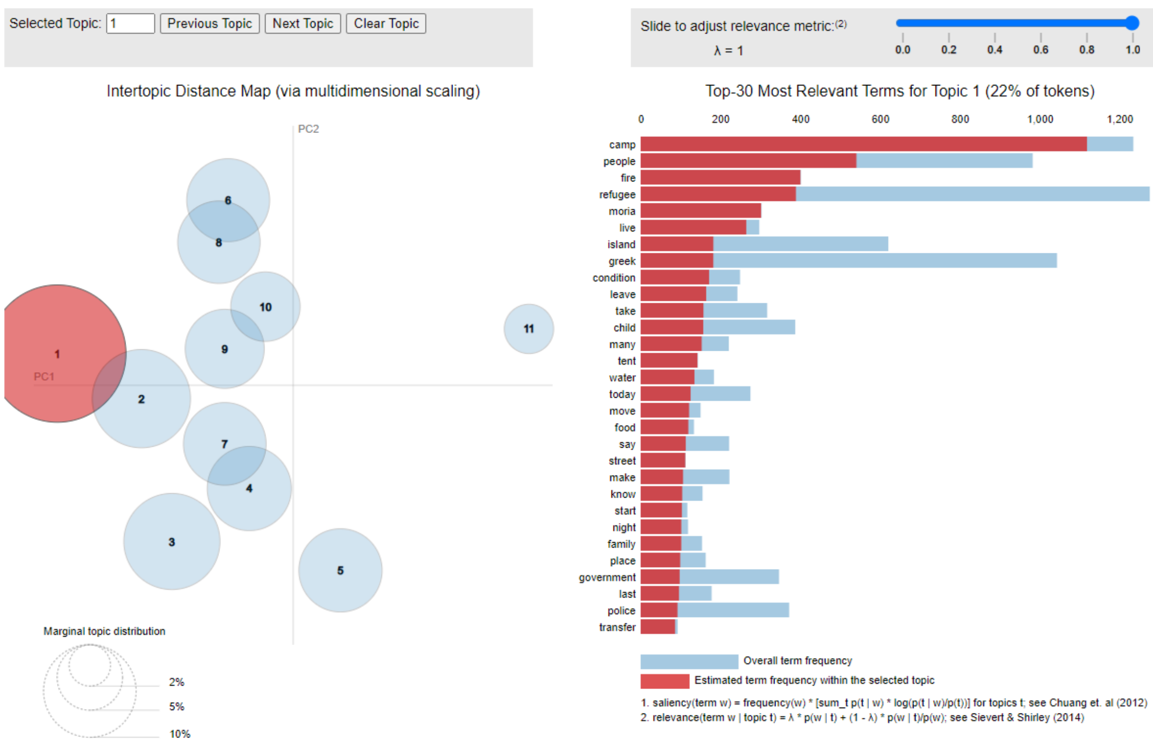

Figure 4-45: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2020

Similar to 2019, tweet volume in 2021 about #RefugeesGR was quite low. This fact also affects the volume of words per topic, so the smaller topics in particular are not as distinctive. Starting with the cluster of topics 1, 2, 3 and 4, all seem to be about the situation on the

102 www.youtube.com/watch?v=O1NNqS62tjw

Aegean islands. Specifically, topics 1 and 2 primarily relate to boat arrivals on the islands, with topic 1 including words such as: “border”, “dead”, “pushback”, “Turkey”, “death”, “miss” and “rescue”, likely referring to shipwrecks and incidents at sea, while topic 2 includes words such as: “local”, “authority”, “group”, “arrive”, “asylum”, “claim”, likely referring to groups of people already having arrived on an island and alerting the local authorities. Topics 3 and 4 relate more to the situation in the camps on the islands, with topic 3 including some prominent words such as “health”, “access”, “mental” and “Samos”, while topic 4 includes words such as “close”, “centre”, “detention” and “violation”. Topic 10 cannot be easily attributed to a single theme, but it’s important to note that it includes a large volume of the words “trial” and “crime” compared to other topics. Topics 5 and 7 are also overlapping with each other and topic 5 mainly refers to conditions in camps, while having a solidarity angle with words such as “need”, “help” and “volunteer”. Topic 7 on the other hand cannot be clearly identified, although it does include words such as “Parwana”, “Amiri” and “Ritsona”, which correspond to a young Afghan refugee, Parwana Amiri[103], who became a published writer[104] while living in Ritsona camp in Greece. Topic 8 seems to refer mainly to requests for donations, with some prominent words being: “support”, “donate”, “monthly”, “donation” and “campaign”. Finally, topic 6, which is the one highlighted in Figure 4-46 below for further reviewing, seems to be solely about pushbacks at sea, with some of the most prominent words being “coast”, “guard”, “Turkey”, “illegal” and “pushback”.

[103] www.opendemocracy.net/en/how-a-young-afghan-woman-trapped-at-europes-borders-found-her-voice
[104] lesvos.w2eu.net/files/2020/04/broshure-Letters-from-Moria-202002-screen.pdf

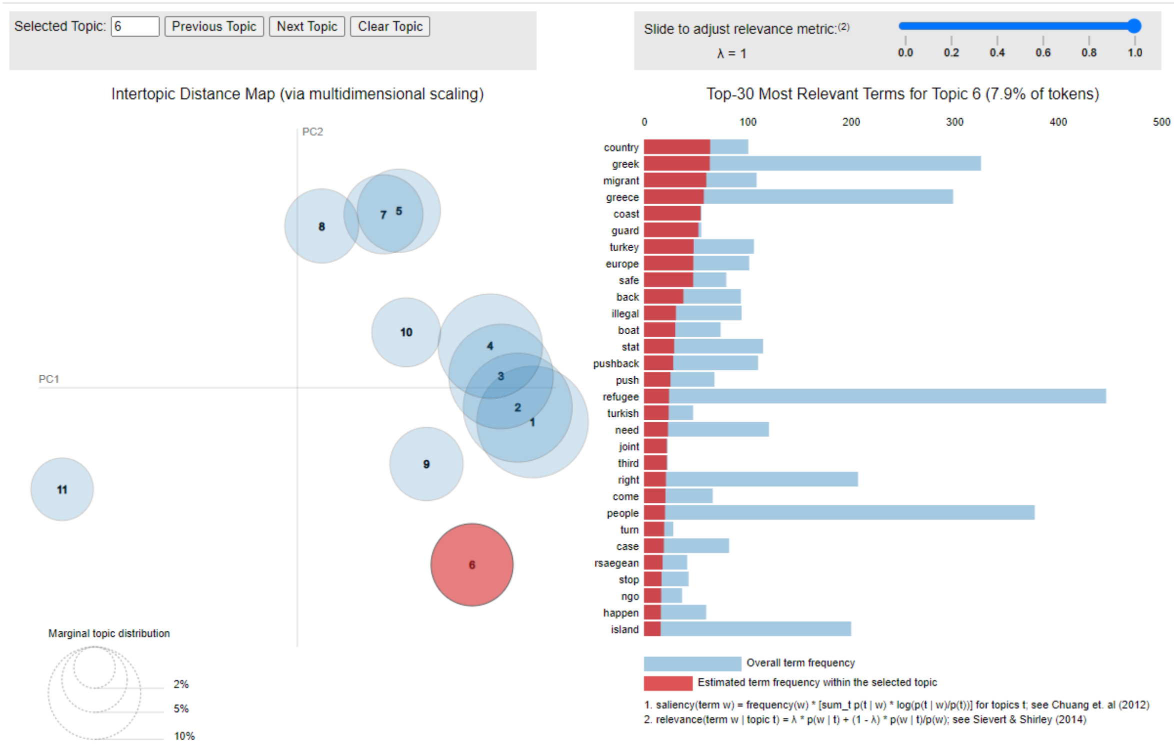

Figure 4-46: LDAvis output for topic modelling of tweet data using the hashtag #RefugeesGR from 2021

4.5 Network Analysis

Progressing into the analysis of relationships between Twitter users involved in the discussion around #RefugeesGR during the period 2015 - 2021, there are two types of interactions that would be of interest within the scope of this dissertation, namely "retweets" (RT) and "mentions" (@). A retweet is a re-posting of a tweet, with or without any additional comments, and it is typically used by Twitter users to promote a certain message or spread some news that are either of interest to them and/or they agree with (Metaxas et al., 2014). A mention on the other hand is the act of tagging another Twitter account to a tweet and it is primarily used to notify another user of a particular piece of information, call another user to action, leave positive or negative feedback for a specific account, or initiate conversation with a user.

Building a network of users based on retweets from the data obtained for this dissertation would not bring significant results, as the data mined from Twitter was based on the tweets that included the hashtag #RefugeesGR and, thus, in order for the algorithm to pull a retweet, the retweet would need to also contain the selected hashtag. Therefore, the most appropriate option here would be to build a network based on users' mentions. The created graph is directed so as to obtain additional measures, such as reciprocity.

Figure 4-47 below presents some popular measures for the graph of Twitter user mentions. There are a total of 10,129 nodes in the Twitter network, corresponding to individual accounts, and 20,104 edges, corresponding to separate connections between those accounts. Moreover, the minimum degree of an account within the Twitter network is 1, meaning that the account is only connected to one other account, while the maximum degree of an account is 675, corresponding to an equal number of connections with other Twitter accounts. However, the most common degree for an account within the Twitter network is 4, which translates to one user account being connected with four other accounts. The density of the Twitter network is almost 0.0002, indicating a sparse network with weaker interconnectivity. Similarly, the reciprocity measure is quite low at approximately 0.02, demonstrating that Twitter users within this dataset are unlikely to mention each other.

```
There are 10129 nodes in the Twitter graph
There are 20104 edges in the Twitter graph
```

```
The maximum degree of a user in the Twitter graph is 675
The minimum degree of a user in the Twitter graph is 1
The average degree of the users in the Twitter graph is 4.0
```

```
The density of the Twitter network is 0.00019597118186453206
The reciprocity of the Twitter network is 0.018404297652208516
```

Figure 4-47: Graph measures for the Twitter mentions network

4.5.1 Network visualization and communities

Once the network of Twitter user mentions is built and its measures are calculated, the graph is imported in the network visualization program, Gephi. The network is visualized in Gephi as an undirected graph, mainly due to the "power imbalance" that inherently exists within this network. This derives from the fact that popular or verified accounts are only mentioning other popular/verified accounts, while regular accounts are also mostly mentioning popular/verified accounts. This approach will only showcase the interaction present between Twitter accounts, without accounting for its orientation.

Figure 4-48 below depicts a macro view of the Gephi visualization of the Twitter user mention network, where the size of the nodes is based on the eigenvector centrality measure. It’s important to note here that the eigenvector centrality measure was chosen to be the one dictating the size of the nodes instead of a simpler degree measure, as it will not merely depict the most well connected nodes, which correspond to the most mentioned accounts, but rather will give an insight on the influence of the nodes within the whole network, taking into consideration nodes that might have fewer connections, but are well connected with very important nodes. Furthermore, the color of the nodes is based on the modularity measure, grouping the nodes together by communities, and the size of the edges is based on their weight measurement, depicting the sum of connections between two nodes.

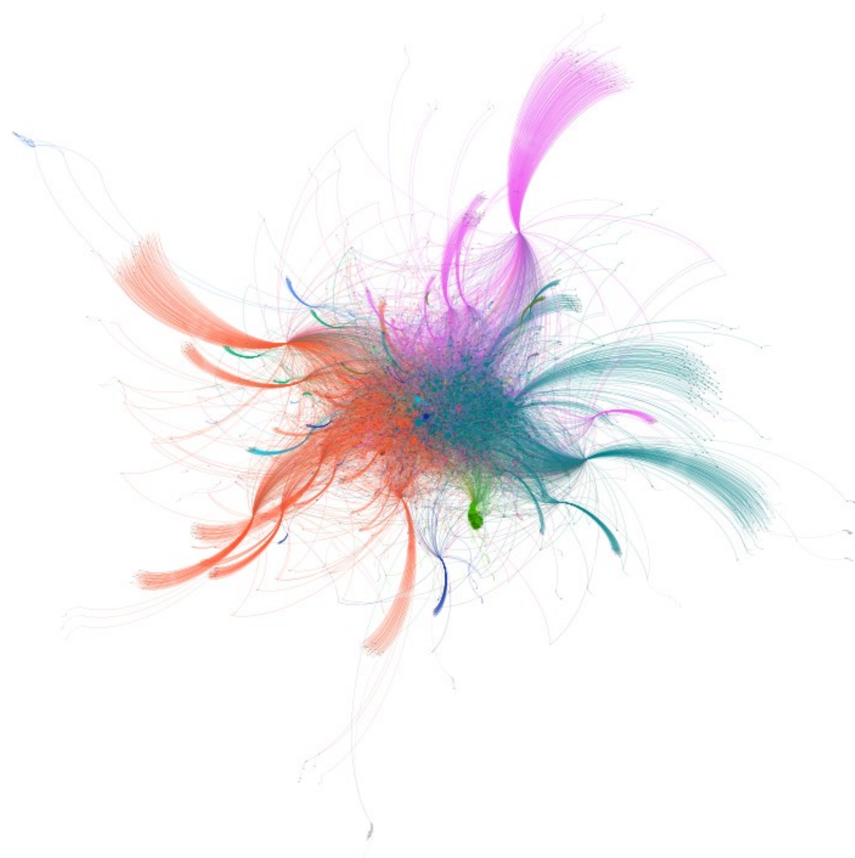

Figure 4-48: Graph of Twitter user mentions on posts about #RefugeesGR

Looking closely at Figure 4-48, the color separation reveals several distinct communities that are detected within the network of Twitter user mentions. Two of those distinct color-coded communities have a widespread reach within the network, while three smaller communities also seem to portray interesting connectivity patterns. Overall, there are five larger communities (with more than 400 nodes each) chosen to be analyzed further, which altogether correspond to 65% of the total network of Twitter user mentions. These five communities are:

1. Teal community, with a percentage coverage of 22%
2. Orange community, with a percentage coverage of 20%
3. Pink community, with a percentage coverage of 13%
4. Green community, with a percentage coverage of 6%
5. Dark Blue community, with a percentage coverage of 4%

The nodes (i.e., Twitter accounts), which comprise the communities are grouped together because the flow of information is persistent among them. This translates to communities of accounts that are frequently mentioned together in tweets regarding #RefugeesGR. While exploring these five communities mentioned above further, an effort will be made to understand and categorize the types of accounts that are grouped together in each community, by highlighting the most important nodes in terms of eigenvector centrality. As the overall number of nodes within each of these five communities ranges from about 400 to 2,000 nodes, focusing mainly on the Hubs, which are the central users within the communities, will provide an overview of the network, without exploring its entirety. At the same time, this methodology omits the accounts amongst these communities which belong to private individuals, who do not seem to be part of a more public lifestyle, which will be protected and not mentioned in this dissertation.

Starting with the teal community, this is the largest grouping within this Twitter network, representing a total of 22% of the whole network. The teal community includes the account that started the hashtag #Regugees_GR, namely @refugees_gr. This community also includes the official accounts of multiple UN organizations, international agencies, and humanitarian organizations, such as UNHCR Greece (@unhcrgreece), Médecins Sans Frontières International (@msf), Médecins Sans Frontières Europe (@msf_sea), Médecins

Sans Frontières Greece (@msfgreece), Amnesty International (@amnesty) and IOM Greece (@iomgreece). Some of the larger nodes within the teal community correspond to accounts of political figures, such as the accounts of the former European Commissioner for Migration, Home Affairs and Citizenship, Dimitris Avramopoulos (@avramopoulos), the Former Minister of Migration Policy, Ioannis Mouzalas (@imouzalas), the account of the President of the European Council (@eucopresident), as well as the now-deactivated international account of the former Prime Minister of Greece, Alexis Tsipras (@tsipras_eu). Only a few Greek civil authority accounts are included within the teal community, namely the account of the Hellenic Police (@hellenicpolice) and a now-deactivated account, which seemingly belonged to the Greek Asylum Service (@greekasylum). Moreover, this community includes several key nodes which correspond to civil society movements that were first formed and became very active during the early stages of the refugee crisis in Greece, such as @areyousyrious, @sol2refugees, @notara26, @noborderkitchen and @pedioareos, and also individual civilian activists who were taking action on the ground helping refugees, whose accounts will be omitted from mentioning here. Although most media outlets are not associated with the teal community, some are included, corresponding to crowdsourced media outlets, such as EFSYN (@efsyntakton) and The Press Project (@thepressproject), the former Official account of ERT, the state-owned public radio and television broadcaster of Greece (@ERTsocial), as well as personal accounts of individual journalists, such as Asteris Masouras (@asteris) and Marianna Karakoulaki (@Faloulah). Figure 4-49 below provides a closer visualization of the teal community and its most important nodes.

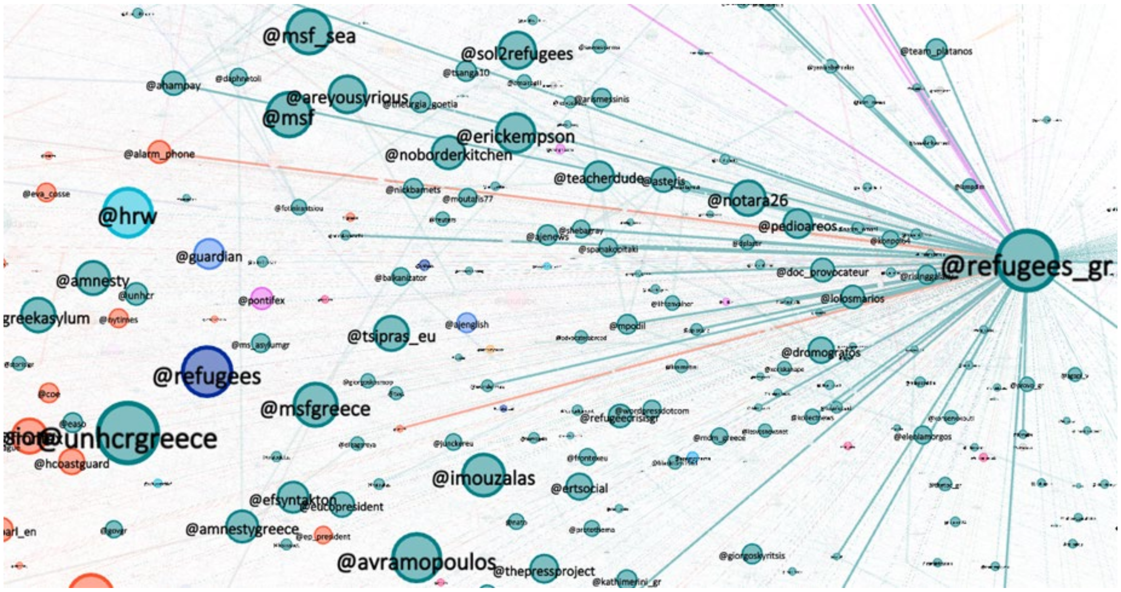

Figure 4-49: Visualization of the largest community within the Twitter user mentions network

Moving on to the orange community, which is the second largest one within this network and includes a significant 20% of the nodes of Twitter user mentions. Many of the most influential accounts within the orange community belong to European Union bodies and representatives, such as the official European Commission account (@eu_commission), the personal accounts of the President and Vice- President of the European Commission, Ursula von der Leyen (@vonderleyen) and Margaritis Schinas (@margschinas), the official account of the President of the European Parliament (@ep_president), the official European Parliament account (@europarl_en), the EU Council account (@eucouncil), the account of the Council of Europe (@coe), the account of the European Border and Coast Guard Agency (@frontex) and the personal account of the EU Commissioner for Home Affairs, Ylva Johansson, (@ylvajohansson). In addition, the orange community also includes nodes corresponding to Greek political bodies, representatives, parties and politicians, such as the official account of the Prime Minister of Greece (@primeministergr), the official account of the Ministry of Migration & Asylum (@migrationgovgr), the accounts of the political parties of Nea Dimokratia (@neademokratia) and SYRIZA (@syriza_gr), as well as the personal accounts of the current Prime Minister of Greece, Kyriakos Mitsotakis (@kmitsotakis), the current Minister of Migration and Asylum, Notis Mitarachi (@nmitarakis) and the former

Minister of Citizen Protection, Michalis Chrysochoides (@chrisochoidis). Within this community, Greek civil authorities are mainly represented by the official account of the Hellenic Coastguard (@hcoastguard), while media outlets primarily include the Athens News Agency - Macedonian Press Agency (@amna_news). The orange community also includes some non-governmental organizations, non-profits and human-right-focused networks active in Greece, such as the Hellenic Union for Human Rights (@hellenic_league), the Greek Forum of Refugees (@refugeegr), the European Council on Refugees and Exiles (@ecre), the Greek Council for Refugees (@gcrefugees), Mobile Info Team (@mobileinfoteam), Refugee Support Aegean (@rspaegean), and Alarm Phone (@alarm_phone). A more detailed visualization of the orange community and its most important nodes can be found in Figure 4-50 below.

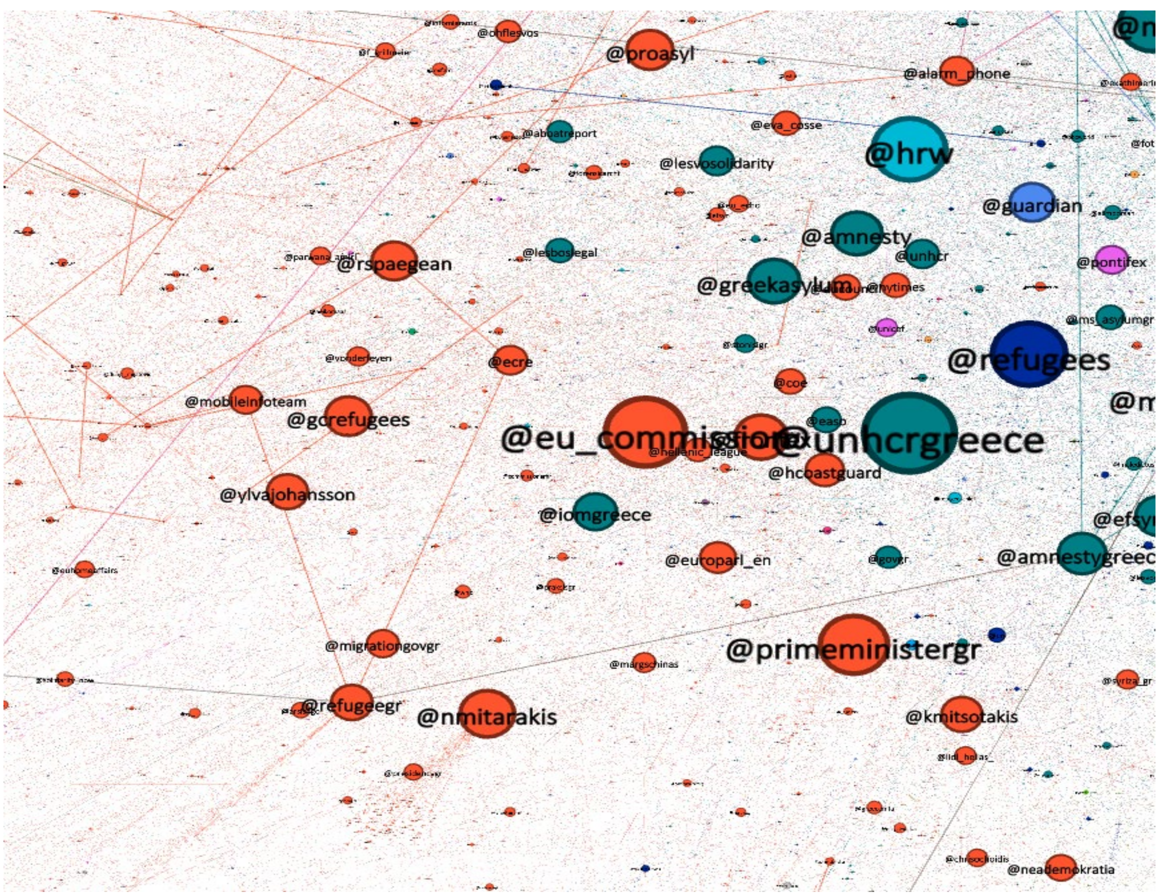

Figure 4-50: Visualization of the second largest community within the Twitter user mentions network

The third largest community within our Twitter mentions network is the pink community, including some 13% of the total number of nodes. The most influential node within this community is the official Twitter account of His Holiness Pope Francis (@pontifex). The official accounts of UNICEF (@unicef) and the Hellenic Red Cross (@hrc_samarites) are also represented within the pink community. Moreover, this community includes Spanish civil society actors and NGOs, such as SOS Refugiados Europa (@sosrefugiados) and Refugee Care (@refugee_care), as well as sea rescue organizations active in the Mediterranean, such as Spanish Open Arms (@openarms_fund) and Sea-Watch (@seawatchcrew). Finally, a large part of this community is comprised of official accounts of foreign news organizations, such as Der Spiegel (@spiegelonline), WELT (@welt), the English account of Radio France Internationale (@rfi_english), N-TV (@ntvde) and Photomovimiento (@fotomovimiento), as well as many personal accounts of journalists, photojournalists, and international media correspondents, such as Patrick Kingsley (@patrickkingsley), Liana Spyropoulou (@lspyropoulou), Santi Palacios (@santipalacios), Matina Stevis-Gridneff (@matinastevis), Zeina Khodr (@zeinakhodraljaz), Yolanda Visser (@assenpoester), Gael Michaud (@pantval), Björn Kietzmann (@bjokie) and Bulent Kilic (@kilicbil). Last but not least, the Twitter account of Chinese artist and activist Ai Weiwei (@aiww) is also included within this community.

Moving on to the green community, which includes only some 6% of the nodes of the network of Twitter user mentions, the most central node is the account of the former Prime Minister of Greece, Alexis Tsipras (@atsipras). Two more accounts of politicians are also included within this community, namely the account of the former President of the European Parliament, Martin Schulz (@martinschulz) and the account of the former Finance Minister of Greece, Yanis Varoufakis (@yanisvaroufakis). Additionally, the green community includes the official Twitter account of the Avaaz Petition platform (@avaaz) and the official account of the Ecumenical Patriarch of the Orthodox Christian Church (@ecupatriarch). Most remaining accounts within this community primarily belong to private individuals, thus their account names and identities will be omitted from listing.

The last community to be discussed here is the dark blue community, which encompasses some 4% of the total number of nodes. The most influential nodes within this community include the official accounts of UNHCR, the UN Refugee Agency (@refugees) and United Nations (@un). Additionally, a few accounts within the dark blue community belong to International NGOs, such as the American Red Cross (@redcross), Amnesty International in the UK (@amnestyuk), Amnesty International USA (@amnestyusa) and Refugees International (@refugeesintl). Yahoo's official Twitter account (@yahoo) is also found within this community, as is the account of the Embassy of the Kingdom of the Netherlands in Greece (@nlingreece). The dark blue community also includes accounts of political parties, such as the account of SYRIZA at the EU Parliament (@syrizaep), and politicians, such as the Vice President of the European Parliament, Dimitris Papadimoulis (@papadimoulis) and the Deputy Minister of National Defence, Nikos Hardalias (@nhardalias), as well as foreign political figures such as the Minister of Foreign Affairs of the Republic of Türkiye, Mevlüt Çavuşoğlu (@mevlutcavusoglu) and former President of the United States of America, Donald Trump (@realdonaldtrump). News organizations are also present within this community, represented by the main account of Radio France Internationale (@rfi), the National Public Radio (@npr), Politico Europe (@politicoeurope), The Economist (@theeconomist), but also including local media outlets, such as Documento (@documentonews) and Kouti Pandoras (@kouti_pandoras). Finally, two accounts of individual journalists are prominent within this community, namely Dalal Mawad (@dalalmawad) and Aida Ghajar (@aidaghajar).

Following the analysis of the most central Hubs of the Twitter mentions network within their specific communities and subsequent classification according to their sectoral role and type of account, it will be useful to also look at this information through a macro lens. For this investigation, the overall amount of mentions each account has accumulated will be considered, along with their inter and intra-related positions within the network, while accounts that act as bridges for subcommunities will also be highlighted. The two most mentioned accounts in this dataset are those of the former Prime Minister of Greece, Alexis Tsipras (@atsipras), and the former President of the European Parliament, Martin Schulz (@martinschulz). Both accounts were grouped within the green community and both act as bridges for their community, as the majority of the green cluster of nodes at the bottom of our network would be disconnected from the rest if those two nodes were not present. Looking at Figure 4-48 above, the green community is located further from the central position of the network, reflecting more niche behavior in comparison to the other large communities. This observation falls in line with the fact that our data spans over a period of sever years and these accounts were more relevant to the hashtag #RefugeesGR when the politicians behind them held central public service positions. The teal community encompasses many of the most mentioned accounts, such as the account that started the hashtag #Regugees_GR (@refugees_gr), the accounts of UNHCR Greece (@unhcrgreece), Médecins Sans Frontières International (@msf) and Médecins Sans Frontières Europe (@msf_sea), which are all very centrally located within the network. Since the Gephi visualization in Figure 4-48 is modelled using a mixture of gravity forces (repulsion and attraction), nodes that are strongly connected attract each other, despite being in different communities. One of the most mentioned accounts from the teal community is the personal account of the journalist Marianna Karakoulaki (@Faloulah), located at the top right edge of the main cluster of the teal community, which also seems to act as a bridge with the teal community nodes that are located at the top right part of the network. Several more nodes from the teal community act as bridges to smaller clusters to the right of the network, however these are accounts of private individuals, and thus will not be highlighted. Continuing with the most mentioned accounts of this dataset, which are also categorized in the orange community, are the personal account of the current Minister of Migration and Asylum, Notis Mitarachi (@nmitarakis), the official account of the European Commission (@eu_commission) and the official account of the Prime Minister of Greece (@primeministergr). These three accounts, and particularly the account of the European Commission, are located centrally within the network, very close to the account of UNHCR Greece from the teal community. Similar to the teal community, there are a few accounts within the orange community that act as bridges to small clusters of nodes, mainly expanding to the left part of the visualization as seen in Figure 4-48. Although most are accounts of private individuals, the account of the Athens News Agency - Macedonian Press Agency (@amna_news), located at the bottom part of the visualization, and the account of the civil society movement Seebrücke Frankfurt (@SeebrueckeFfm), located at the top left of the visualization, stand out. ANA-MPA node acts as a bridge, connecting the network to an orange cluster of nodes at the bottom of the visualization, which belong mainly to politicians and other media outlets, while Seebrücke node connects the network to an orange cluster on the top left of the visualization, which mainly belong to other civil society groups, press correspondents and activists. The last two highly mentioned nodes that will be highlighted here are the main account of UNHCR, the UN Refugee Agency (@refugees), grouped with the dark blue community, and the account of the international NGO Human Rights Watch (@hrw), grouped with the light blue community. Those two important hubs are also located at the center of the visualization in Figure 4-48, approximately where the teal and orange communities meet, and their position is a direct result of the multiple connections of both nodes to these two communities.

Concluding with the presentation of findings from the analysis of the dataset of tweets mined from Twitter, some key takeaways will be presented in the discussion Chapter 5.

5 Discussion

This dissertation offers a testimony on the increasingly relevant role of social media, and Twitter in particular, within the public discourse. Choosing a hashtag, namely #RefugeesGR, that was quite central to the refugee discussion in Greece during the period 2015 – 2021 and was used by a wide range of local and international actors, proved crucial for mining historical tweets that captured a large aspect of the conversation about refugees.

The hypothesis that Twitter as a source is suitable to offer insight on real world events in a systematic manner can be largely accepted, albeit within the limitations of selection bias arising from mining large datasets, discussed later in this chapter. The extensive analysis of events that sparked the most traffic on Twitter around #RefugeesGR brought forth some important developments of the refugee crisis in Greece during the period 2015 – 2021, more so for the years with high volume of tweets overall. The majority of these high-traffic events seem to be focused on incidents involving hardship, struggle and clashes, which could be somewhat attributed to the fact that media outlets do disproportionately cover violent events (O'Hear, 2020).

The findings in this work speak to the powerful tool of topic analysis, as LDA proved to be quite effective in producing a few distinct themes throughout the yearly analysis, despite the complexity of the subject matter that has innately intertwined themes revolving around the lives of refugees in Greece. Topic modelling produced better results and more distinct clusters for years with high tweet volumes, similar to event analysis, but its limitations mainly affected smaller topics. Reviewing the results of topic analysis on a yearly basis provided insight as to which topics maintained their popularity within the refugee discourse throughout the period 2015 – 2021, which was also in alignment with the recurrence of popular hashtags and most used words per year. The persistent themes throughout the years revolved around discussion of new boat arrivals on the islands, the living conditions in formal and informal camps across Greece and calls-to-action for humanitarian aid and in solidarity with refugees. The topic analysis also revealed a shift in the geographic interest regarding refugees in Greece, from the border region with North Macedonia during 2015 and 2016 to the border region with Turkey and the Aegean islands from 2017 onwards. This result is in line with the real on-the-ground situation in Greece at the time, described in detail in Chapter 4.2, deriving from the operation of Idomeni camp at the border between Greece and North Macedonia during 2015 – 2016 and the focus on the EU-Turkey deal, along with events at the border region in Evros and the Aegean islands from 2017 onwards.

Turning to Twitter accounts utilizing the hashtag #RefugeesGR, the most influential accounts emerging from the analysis belonged to International Humanitarian Organizations, EU governmental bodies, and individual (international and local) politicians. Building a network based on Twitter user interactions (specifically mentions) between professional, verified or popular accounts and smaller personal accounts, the inherent "power imbalance" was taken into account and, thus, it was imperative to use eigenvector centrality as a measure of visualizing the network instead of simple degree centrality. Community analysis formed five major clusters, corresponding to some 65% of the total network of Twitter user mentions, which were further explored by grouping the types of user accounts within those five communities. Perhaps unsurprisingly, accounts were grouped together not based on similar characteristics of the nodes in terms of user type (i.e., accounts of media outlets or governmental bodies), but rather on the nodes that were co-serving a particular goal. This community grouping shed a nuanced light on how Twitter users bring more powerful accounts together, especially when it comes to call-to-action tweets about the refugee hardships in Greece. This mechanism resulted in UN Organizations, International Humanitarian Agencies, EU governmental bodies and international media outlets being grouped together with particular political figures and seems to be a consequence of the different time-periods when those political figures were relevant to the hashtag #RefugeesGR, perhaps due to holding a public service position during the span of the seven years examined in this work. A few local media outlets, individual journalists and activist groups also played crucial roles within the network and their specific communities, acting as bridges to smaller niche clusters of users on the outer parts of the network.

All in all, this dissertation's findings can attest to the value of tweets as a good basis for data mining that has potential to yield real-world insights regarding the public discourse of a given topic and the type of Twitter user accounts that drive it.

Diving into any large dataset, especially one mined from social media, requires attention to potential limitations around sample selection and sampling bias, as big datasets, despite their large number of participants and variables, may entail issues of population underrepresentation (Ashton et al., 2016). This observation, coupled with the fact that Twitter penetration rates vary greatly between countries and age groups should be taken into consideration when utilizing information from these sources. Moreover, it's important to reiterate that the body of work within this dissertation is directly correlated with a specific hashtag, namely #RefugeesGR, and its use on Twitter over the period 2015 – 2021, which inevitably does not encompass the entirety of the conversation on Twitter regarding the refugee crisis in Greece. Finally, processing and analyzing the topics of tweets written in English allows this work to present a more globalized view of the refugee situation in Greece that is, however, not necessarily identical to the one that would have been presented if the Greek tweets were selected for analysis. The challenges of the work presented in this dissertation relate mainly to the retrieval of tweets from Twitter, as presented in Chapter 3.1 on data mining, and should be addressed for any future research. On 2 February 2023, Twitter announced the end of its free API access[105], launching a paid version instead, which will likely impact all data mining options that focus on Twitter data. Deleted and suspended Twitter accounts portrayed another important impediment during the analysis phase of this dissertation, as, although the deleted accounts' past activity could be retrieved, it was no longer possible to get a real-time overview and wholly evaluate these accounts beyond a few tweets relevant to this work.

The work in this dissertation could be further expanded by utilizing the acquired dataset in different ways. A methodology that could potentially offer further insight into the discussion on #RefugeesGR would be to perform topic analysis, coupled with analysis of the most highlighted events over the period 2015 – 2021, by using only tweets from a specific sub-community of users, namely media outlets, international humanitarian organizations, political figures and private accounts, to understand the differences in language used and events promoted publicly between these communities. The topic modelling component of the work herein could also benefit by presenting an alternative method to LDA, and specifically

[105] https://twitter.com/TwitterDev/status/1621026986784337922

Non-Negative Matrix Factorization (NMF), as it is argued to be more efficient for shorter texts. Additionally, the network analysis part of the work presented here would probably yield different patterns if the connections between the users were analyzed on the basis of smaller time periods (e.g., periods corresponding to different government schemes in Greece). Finally, this dataset could be further utilized through sentiment analysis, perhaps separately for English and Greek tweets for comparative investigations.

References

Ashton, W., Bhattacharyya, P., Galatsanou, E., Ogoe, S., & Wilkinson, L. (2016). *Emerging Uses of Big Data in Immigration Research*. 22.

Aswad, F., Hamoodat, H., Ribeiro, E., & Menezes, R. (2020). Communities of Human Migration in Social Media: An Experiment in Social Sensing. In H. Barbosa, J. Gomez-Gardenes, B. Gonçalves, G. Mangioni, R. Menezes, & M. Oliveira (Eds.), *Complex Networks XI* (pp. 222–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-40943-2_19

Aswad, F., & Menezes, R. (2018, May 23). *Refugee and Immigration: Twitter as a Proxy for Reality*.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–362.

Bermingham, A., & Smeaton, A. F. (2010, October). *Classifying sentiment in microblogs: Is brevity an advantage?* CIKM 2010 - 19th International Conference on Information and Knowledge Management, Toronto, Canada. http://dx.doi.org/10.1145/1871437.1871741

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). *Reading Tea Leaves: How Humans Interpret Topic Models*. 9.

Freire, Y., & Graells-Garrido, E. (2019). Characterization of Local Attitudes Toward Immigration Using Social Media. *Companion Proceedings of The 2019 World Wide Web Conference*, 783–790. https://doi.org/10.1145/3308560.3316455

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hübl, F., Cvetojevic, S., Hochmair, H., & Paulus, G. (2017). Analyzing Refugee Migration Patterns Using Geo-tagged Tweets. *ISPRS International Journal of Geo-Information*, *6*(10), 302. https://doi.org/10.3390/ijgi6100302

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PloS One*, *9*, e98679. https://doi.org/10.1371/journal.pone.0098679

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, *20*(3), 531–558. https://doi.org/10.1007/s10796-017-9810-y

Lafferty, J. D., & Blei, D. M. (2009). Topic models. In *Text Mining: Classification, Clustering, and Applications* (pp. 71–94). Chapman and Hall/CRC.

Laucuka, A. (2018). Communicative Functions of Hashtags. *Economics and Culture*, *15*, 56–62. https://doi.org/10.2478/jec-2018-0006

Marres, N., & Weltevrede, E. (2013). Scraping the Social? *Journal of Cultural Economy*, *6*(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). *OpenOrd: An open-source toolbox for large graph layout*. 786806. https://doi.org/10.1117/12.871402

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Meraz, S., & Papacharissi, Z. (2013). Networked Gatekeeping and Networked Framing on #Egypt. *The International Journal of Press/Politics*, *18*(2), 138–166. https://doi.org/10.1177/1940161212474472

Metaxas, P., Mustafaraj, E., Kily, W., Zeng, L., O'Keefe, M., & Finn, S. (2014). *Do Retweets indicate Interest, Trust, Agreement? (Extended Abstract)*.

Mifrah, S. (2020). Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 5756–5761. https://doi.org/10.30534/ijatcse/2020/231942020

Nerghes, A., & Lee, J.-S. (2018). The Refugee/Migrant Crisis Dichotomy on Twitter: A Network and Sentiment Perspective. *Proceedings of the 10th ACM Conference on Web Science*, 271–280. https://doi.org/10.1145/3201064.3201087

Nerghes, A., & Lee, J.-S. (2019). Narratives of the Refugee Crisis: A Comparative Study of Mainstream-Media and Twitter. *Media and Communication*, *7*(2), 275–288. https://doi.org/10.17645/mac.v7i2.1983

O'Hear, M. (2020). Violent Crime And Media Coverage In One City: A Statistical Snapshot. *MARQUETTE LAW REVIEW*.

Reyes-Menendez, A., Saura, J., & Alvarez-Alonso, C. (2018). Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach. *International Journal of Environmental Research and Public Health*, *15*(11), 2537. https://doi.org/10.3390/ijerph15112537

Rogers, R. (2013). Debanalizing Twitter: The transformation of an object of study. *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, 356–365. https://doi.org/10.1145/2464464.2464511

Siapera, E., Boudourides, M., Lenis, S., & Suiter, J. (2018). Refugees and Network Publics on Twitter: Networked Framing, Affect, and Capture. *Social Media + Society*, *4*, 205630511876443. https://doi.org/10.1177/2056305118764437

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/W14-3110

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I., Pappalardo, L., Passarella, A., Pedreschi, D., Pollacci, L., Pratesi, F., & Sharma, R. (2021). Human migration: The big data perspective. *International Journal of Data Science and Analytics*, *11*(4), 341–360. https://doi.org/10.1007/s41060-020-00213-5

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). *Exploring Topic Coherence over Many Models and Many Topics*. 11.

Vázquez, M. F., & Pérez, F. S. (2019). Hate Speech in Spain Against Aquarius Refugees 2018 in Twitter. *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, 906–910. https://doi.org/10.1145/3362789.3362849

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815478

Wetherell, C., Plakans, A., & Wellman, B. (1994). Social Networks, Kinship, and Community in Eastern Europe. *The Journal of Interdisciplinary History*, *24*(4), 639–663. https://doi.org/10.2307/205629

Appendix

```
#Install as needed
#!pip install git+https://github.com/JustAnotherArchivist/snscrape.git
#!pip install wordcloud
#!pip install textblob
#!pip install plotly
#!pip install -U gensim
#nltk.download("stopwords")
#pip install pyldavis

#Import packages
import os
import glob
import pandas as pd
import json
from pandas.io.json import json_normalize
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
import numpy as np
import plotly.express as px
import plotly.graph_objs as go
import seaborn as sns
import re
import spacy
import nltk
from nltk.corpus import stopwords
from nltk import bigrams
import string
import itertools
import collections
from wordcloud import WordCloud, STOPWORDS
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel
import logging
from __future__ import print_function
import pyLDAvis
import pyLDAvis.sklearn
import pyLDAvis.gensim_models
pyLDAvis.enable_notebook()
nlp = spacy.load("en_core_web_sm")
import csv
from operator import itemgetter
import networkx as nx
from networkx.algorithms import community
```

```python
#Scraping Twitter for tweets that include the Hashtag #RefugeesGR from 1
January 2015 to 31 December 2021 using an Snscrape text-based query
#Setting variables for the command string
max_tweets = 300000
text = "#refugeesGR"
since_date = "2015-01-01"
until_date = "2021-12-31"


#Read the generated json file from the CLI command and create a pandas
dataframe
raw_tweets = pd.read_json('refugee_tweets_2015_2021.json', lines=True)
#Export dataframe into a csv file
raw_tweets.to_csv('refugee_tweets_2015_2021.csv', sep=',', index=False,
encoding='utf-8-sig')


#Normalizing objects from the .json file and creating .csv files with all
the needed data into normalized columns
#Read JSON file containing tweets data
raw_tweets = pd.read_json('./refugee_tweets_2015_2021.json', lines=True)

print("Tweets shape: ", raw_tweets.shape)
raw_tweets.head()

#Normalize 'user' field
users = json_normalize(raw_tweets['user'])
users.drop(['_type', 'description', 'descriptionUrls',
'profileBannerUrl', 'linkTcourl', 'protected', 'profileImageUrl',
'profileImageUrl', 'label', 'label._type', 'label.description',
'label.url', 'label.badgeUrl', 'label.longDescription'], axis=1,
inplace=True)
users.rename(columns={'id':'userId', 'url':'profileUrl',
'rawDescription':'userDescription'}, inplace=True)

#Create DataFrame and remove duplicates of users
users = pd.DataFrame(users)
users.drop_duplicates(subset=['userId'], inplace=True)
print("Users shape: ", users.shape)
users.head()

#Add column for 'userId'
user_id = []
for user in raw_tweets['user']:
    uid = user['id']
    user_id.append(uid)
raw_tweets['userId'] = user_id


#Remove less important and duplicate columns
raw_tweets.drop(['_type', 'content', 'user', 'conversationId',
'sourceUrl', 'sourceLabel', 'outlinks', 'tcooutlinks',
'inReplyToTweetId', 'inReplyToUser', 'cashtags'], axis=1, inplace=True)
```

```
raw_tweets.rename(columns={'url':'tweetUrl', 'id':'tweetId', },
inplace=True)

#Convert to DataFrame and remove duplicates
raw_tweets = pd.DataFrame(raw_tweets)
raw_tweets.drop_duplicates(subset=['tweetId'], inplace=True)
print("Tweets shape: ", raw_tweets.shape)

#Merge the two dataframes "raw_tweets" and "users" into one, based on the
'userId'
merged_dfs = pd.merge(raw_tweets, users, on='userId')
merged_dfs.to_csv("refugeesGR_2015_2021.csv", index = False,
encoding='utf-8-sig')
print("Merged file shape: ", merged_dfs.shape)
merged_dfs.head()


#Split csv to two files containing only English and Greek tweets
respectively
file = pd.read_csv("refugeesGR_2015_2021.csv")

Greek = file[file['lang'] == 'el']
English = file[file['lang'] == 'en']

Greek.to_csv('refugeesGR_2015_2021_GR.csv', index=False, encoding='utf-8-
sig')
English.to_csv('refugeesGR_2015_2021_ENG.csv', index=False,
encoding='utf-8-sig')



#Plotting the volume of tweets in the dataset by language used
#Read data files
tweets = pd.read_csv("./refugeesGR_2015_2021.csv", encoding='utf-8-sig')
tweetsGR = pd.read_csv("./refugeesGR_2015_2021_GR.csv", encoding='utf-8-
sig')
tweetsENG = pd.read_csv("./refugeesGR_2015_2021_ENG.csv", encoding='utf-
8-sig')

#Adding a new column, holding information on the month and year each
tweet was created
tweets['month_year'] = pd.to_datetime(tweets['date']).dt.to_period('M')
tweetsGR['month_year'] =
pd.to_datetime(tweetsGR['date']).dt.to_period('M')
tweetsENG['month_year'] =
pd.to_datetime(tweetsENG['date']).dt.to_period('M')



#Grouping tweets by month, using the new column
tweets_by_month =
pd.to_datetime(tweets['date']).dt.to_period('M').value_counts().sort_inde
x()
```

```
tweets_by_month.index = pd.PeriodIndex(tweets_by_month.index)
tweets_per_month =
tweets_by_month.rename_axis('month').reset_index(name='tweets')

#Grouping Greek tweets by month
GRtweets_by_month =
pd.to_datetime(tweetsGR['date']).dt.to_period('M').value_counts().sort_in
dex()
GRtweets_by_month.index = pd.PeriodIndex(GRtweets_by_month.index)
GRtweets_per_month =
GRtweets_by_month.rename_axis('month').reset_index(name='GRtweets')
GRtweets_per_month

#Grouping English tweets by month
ENGtweets_by_month =
pd.to_datetime(tweetsENG['date']).dt.to_period('M').value_counts().sort_i
ndex()
ENGtweets_by_month.index = pd.PeriodIndex(ENGtweets_by_month.index)
ENGtweets_per_month =
ENGtweets_by_month.rename_axis('month').reset_index(name='ENGtweets')
ENGtweets_per_month

#Merging the dataframes with the summary data on all three tweet sets
(all tweets, Greek tweets and English tweets)
mergeGR = pd.merge(tweets_per_month, GRtweets_per_month, on='month')
mergeGRENG = pd.merge(mergeGR, ENGtweets_per_month, on='month')

#Plotting all three summary information on tweets per month (for all
tweets, Greek tweets and English tweets)
fig = go.Figure()

fig.add_trace(go.Scatter(x=mergeGRENG['month'].astype(dtype=str),
                        y=mergeGRENG['tweets'],
                        marker_color='navy',
                        name='all tweets'))
fig.add_trace(go.Scatter(x=mergeGRENG['month'].astype(dtype=str),
                        y=mergeGRENG['GRtweets'],
                        marker_color='firebrick',
                        name='Greek tweets'))
fig.add_trace(go.Scatter(x=mergeGRENG['month'].astype(dtype=str),
                        y=mergeGRENG['ENGtweets'],
                        name='English tweets',
                        marker_color='cadetblue'))
fig.update_layout({"title": 'Tweets with Hashtag #RefugeesGR from April
2015 to December 2021',
                   "xaxis": {"title":"Timeline"},
                   "yaxis": {"title":"Total tweets"}})

fig.show()

#Plot Hashtags
data = pd.read_csv("./refugeesGR_2015_2021_ENG.csv")
```

```
#Extract hashtags from tweets
def hashtag_extraction(x):
    hashtags = []
    #Loop over the words in each tweet to find Hashtags
    for i in x:
        ht = re.findall(r"#(\w+)", i)
        hashtags.append(ht)
    return hashtags

#Format dates and create a new column just for the year
data['date'] = data['date'].astype('datetime64[ns]')
data['Year'] = data['date'].apply(lambda x: "%d" % (x.year))

#Extract Hashtags from English tweets per Year
Hashtags_all = hashtag_extraction(data['renderedContent'])
Hashtags_all = sum(Hashtags_all,[])
Hashtags_all = np.char.lower(Hashtags_all)

#Plot the 20 most popular Hashtags for the whole dataset
a = nltk.FreqDist(Hashtags_all)
d = pd.DataFrame({'Hashtag': list(a.keys()),
                  'Count': list(a.values())})

#Select top 20 most frequent hashtags
d = d.nlargest(columns = "Count", n = 20)
plt.figure(figsize = (15, 7))
ax = sns.barplot(data = d, x = "Count", y = "Hashtag")
ax.set(ylabel = 'Hashtags', xlabel = 'Volume of Hashtags', title = "Most
popular Hashtags used with #RefugeesGR")
plt.savefig('Hashtags.png')
plt.show()



#Preprocessing tweets
#Read data file
tweetsENG = pd.read_csv("./refugeesGR_2015_2021_ENG.csv")
#Stopwords
stop_words = stopwords.words('english')
#Add collection and interference words to stopwords
stop_words.extend(['refugeesgr', 'youtu', 'youtube', 'twitter',
'instagram', 'facebook', 'blogspot', 'blog', 'policevoice', 'utw',
'theguardian', 'provocateu', 'doc', 'ekathimerini', 'global',
                   'insight', 'amna', 'article', 'tweet'])

#Remove Hashtags, numerals, mentions, links and small interference words
from tweets
def remove_noise(tweet):
    tweet = re.sub(r'\b(\w{1,3})\b', '', tweet)
    tweet = re.sub(r"http\S+", '', tweet)
    tweet = re.sub(r'\S+\.com\S+','', tweet)
    tweet = re.sub(r'\@\w+', '', tweet)
    tweet = re.sub(r'\#\w+', '', tweet)
```

```
    tweet = re.sub('[^A-Za-z]', ' ', tweet.lower())
    return tweet

#Run database through the function for removing noise
tweetsENG['tweets'] = tweetsENG['renderedContent'].apply(lambda x:
remove_noise(x))
#Saving cleaned tweet database to a csv file
tweetsENG.to_csv('refugeesGR_2015_2021_ENG_clean.csv', index=False,
encoding='utf-8-sig')



#Cleaning tweets
#Fix date formatting and add a column specifying the year of each tweet
tweetsENG['date'] = tweetsENG['date'].astype('datetime64[ns]')
tweetsENG['year'] = tweetsENG['date'].apply(lambda x: "%d" % (x.year))

#Remove stopwords and collection words and lemmatize the list of tweet
words
def clean_tweets(tweets, stop_words=stop_words, allowed_postags=['NOUN',
'ADJ', 'VERB', 'ADV']):
    tweets = [[word for word in simple_preprocess(str(txt)) if word not
in stop_words] for txt in tweets]
    output = []
    for text in tweets:
        txt = nlp(" ".join(text))
        output.append([token.lemma_ for token in txt if token.pos_ in
allowed_postags])

    #Remove stopwords once more after lemmatization
    output = [[word for word in simple_preprocess(str(txt)) if word not
in stop_words] for txt in output]
    return output

#Creating a list, tokenizing tweets and running the function to remove
stopwords and lemmatize tweets
tweet_list = tweetsENG['tweets'].tolist()
tweet_words = [tweet.split() for tweet in tweetsENG['tweets']]
tweet_corpus = clean_tweets(tweet_words)
tweet_corpus



#Investigate the optimal number of topics for the corpus
#Calculate coherence scores for different numbers of topics to determine
the optimal number
def calc_coh_val(dictionary, corpus, texts, start, limit, step):
    coherence_scores = []
    model_list = []
    for num_topics in range(start, limit, step):
        model = gensim.models.ldamodel.LdaModel(corpus = dtm,
                                                id2word = dix,
                                                num_topics = num_topics)
        model_list.append(model)
```

```
        coherencemodel = CoherenceModel(model = model,
texts=tweet_corpus, dictionary=dix, coherence='c_v')
        coherence_scores.append(coherencemodel.get_coherence())
    return model_list, coherence_scores

#Creating a dictionary from the tweet corpus
dix = corpora.Dictionary(tweet_corpus)
#Converting the tweet corpus into the Document Term Matrix
dtm = [dix.doc2bow(word) for word in tweet_corpus]
#Calculate coherence scores per model based on the number of topics
model_list, coherence_scores = calc_coh_val(dictionary = dix, corpus =
dtm, texts = tweet_corpus, start = 5, limit = 21, step = 1)

#Plot a graph for the results
start = 5
limit = 21
step = 1
x = range(start, limit, step)
plt.plot(x, coherence_scores)
plt.xlabel("Number of topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_scores"), loc = 'best')
plt.show()

#Print the coherence scores per model
for n, cv in zip(x, coherence_scores):
    print("The model with", n, "topics has a coherence score of",
round(cv, 3))


#Fitting an LDA model on the whole 2015 - 2021 tweet dataset using gensim
library and the optimal number of topics determined above
#Running and training the LDA model on the Document Term Matrix
lda_model = gensim.models.ldamodel.LdaModel(corpus = dtm,
                                           id2word = dix,
                                           num_topics = 10,
                                           random_state = 3,
                                           passes = 23,
                                           iterations = 399,
                                           per_word_topics = True,
                                           alpha='auto')

#Create (and save extrernally) a visualization of the topics using
pyLDAvis
visual = pyLDAvis.gensim_models.prepare(lda_model, dtm, dix)
pyLDAvis.save_html(visual, 'LDA_FullDataset.html')


#Calculate the model's perplexity score
print('\nPerplexity score:', lda_model.log_perplexity(dtm))

#Print the topics
```

```
model_topics = lda_model.show_topics(formatted = True)
lda_model.show_topics(num_topics = 10, num_words = 10, log = False,
formatted = True)


#Wordcloud of the most popular words used in the whole 2015 - 2021 tweet
dataset
fig_all = " ".join(" ".join(x) for x in tweet_corpus)

wordcloud = WordCloud(width = 1300, height = 900,
                background_color ='white', max_words = 100,
min_word_length = 4,
                min_font_size = 10, max_font_size=170).generate(fig_all)

plt.figure(figsize = (9, 9), facecolor = 'black')
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)

plt.savefig('Wordcloud_ALL.png')
print('Most used words in tweets with Hashtag #RefugeesGR from 2015 to
2021')
plt.show()


#Create a second dataframe to put user information
users = pd.DataFrame(columns = ["userID"])

#Populate the user dataframe with important attributes
users["username"] = tweets["username"]
users["userID"] = tweets["userId"]
users["followersCount"] = tweets["followersCount"]
users["verified"] = tweets["verified"]
users["friendsCount"] = tweets["friendsCount"]
users["statusesCount"] = tweets["statusesCount"]

#Remove duplicates
users.drop_duplicates(subset=['userID'], inplace=True)
users.sort_values(by=["followersCount"], ascending=False).head(20)


#Build a bubblemap using scatterplot
plt.rcParams['figure.figsize'] = [11, 9]
sns.scatterplot(data=users, x="followersCount", y="friendsCount",
size="statusesCount", hue="verified", alpha=0.7, legend=True, sizes=(10,
2500))
plt.title('Twitter users using the Hashtag #RefugeesGR by followers and
friend counts')
plt.xlabel('Number of followers')
plt.ylabel('Number of friends')
plt.savefig('TwitterUsersBubblemap.png')
plt.show()
```

```
#Plotting the 20 most active Twitter users
a = nltk.FreqDist(tweets["username"])
d = pd.DataFrame({'Username': list(a.keys()),
                  'Count': list(a.values())})
d = d.nlargest(columns = "Count", n = 20)
plt.figure(figsize = (15, 11))
ax = sns.barplot(data = d, x = "Count", y = "Username", palette="crest")
ax.set(ylabel = 'Username', xlabel = 'Volume of Tweets', title = "Most
active Twitter users using the Hashtag #RefugeesGR")
plt.savefig('MostActiveUsers.png')
plt.show()


#Plotting the 20 most followed Twitter users
plt.figure(figsize = (15, 11))
d = users.nlargest(columns = "followersCount", n = 20)
ax = sns.barplot(data=d, x="followersCount", y="username",
palette="rocket")
ax.set(ylabel = 'Username', xlabel = 'Followers (in millions)', title =
"Most followed Twitter users using the Hashtag #RefugeesGR")
plt.savefig('MostFollowedUsers.png')
plt.show()


#Build a network from user interactions
#Creating a directed graph with networkx
G = nx.DiGraph()

#Iterate tweet dataframe to extract mentioned users
for tweet in tweets.iterrows():
    user = tweet[1]['username']
    user = user.lower()
    user = f'@{user}'
    text = tweet[1]['renderedContent']
    text = text.lower()
    mentions = set(re.findall(r"@(\w+)", text))

    #Add users who are mentioned in the tweets to the graph
    for mention in mentions:
        mention = f'@{mention}'
        G.add_edge(user, mention)

#Save the graph
interactions = nx.to_pandas_edgelist(G)
interactions.to_csv('TwitterInteractions.csv', index=False)

#Show some network metrics
print(f"There are {G.number_of_nodes()} nodes in the Twitter graph")
print(f"There are {G.number_of_edges()} edges in the Twitter graph")
```

```
degrees = [deg for (node, deg) in G.degree()]
print(f"The maximum degree of a user in the Twitter graph is
{np.max(degrees)}")
print(f"The minimum degree of a user in the Twitter graph is
{np.min(degrees)}")
print(f"The average degree of the users in the Twitter graph is
{np.mean(degrees):.1f}")
print(f"The density of the Twitter network is {nx.density(G)}")
print(f"The reciprocity of the Twitter network is {nx.reciprocity(G)}")
print(f"The average clustering coefficient of the Twitter network is
{nx.average_clustering(G)}")
print(f"The transitivity of the Twitter network is {nx.transitivity(G)}")


#Run metrics for eigenvector & betweenness centrality
eigen = nx.eigenvector_centrality(G)
between = nx.betweenness_centrality(G)

#Assign above metrics as attributes to nodes in the network
nx.set_node_attributes(G, eigen, 'eigenvector')
nx.set_node_attributes(G, between, 'betweenness')

#Show top users by eigenvector centrality
eigen_sort = sorted(eigen.items(), key=itemgetter(1), reverse=True)
print("Top 20 Twitter users with the highest eigenvector centrality in
the network:")
for user in eigen_sort[:20]:
    print(user)


#Show top users by betweenness centrality
between_sort = sorted(between.items(), key=itemgetter(1), reverse=True)
print("Top 20 Twitter users with the highest betweenness centrality in
the network:")
for user in between_sort[:20]:
    print(user)


#Detect communities in the Twitter user mentions network
communities = community.greedy_modularity_communities(G)

#Loop through the list of communities to find users that belong in the
same communities
modularity_dict = {}
for i,c in enumerate(communities):
    for name in c:
        modularity_dict[name] = i

#Add modularity attribute to nodes in the network
nx.set_node_attributes(G, modularity_dict, 'modularity')
```

```
#For large communities with more than 400 nodes, print out their members
for i,c in enumerate(communities):
    if len(c) > 400:
        print('Class '+str(i)+':', list(c))


#Find the nodes of the largest community in the network and their
eigenvector centrality measures
class0 = [n for n in G.nodes() if G.nodes[n]['modularity'] == 0]
class0_eigen = {n:G.nodes[n]['eigenvector'] for n in class0}

#Sort by eigenvector centrality
class0_sort = sorted(class0_eigen.items(), key=itemgetter(1),
reverse=True)

print("Modularity Class 0 Sorted by Eigenvector Centrality:")
for node in class0_sort[:30]:
    print("Twitter User:", node[0], "| Eigenvector Centrality:", node[1])


#Export the graph in gephi format to continue the visualization on that
software
nx.write_gexf(G, 'twitter_interactions_network.gexf')
```