Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Investigating the Potential of Topic Modelling On Social Media to Support the Identification of Refugee Needs

Misbah Rizaee

A Dissertation

Submitted to the University of Dublin, Trinity College

In Partial Fulfilment of the Requirements for the Degree of

Master of Science in Computer Science (Future Networked Systems)

Supervisor: Professor Timothy Savage

August 2022

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university.

Misbah Rizaee

University of Dublin, Trinity College

August 19, 2022

Permission to lend and/or copy

I agree that the Trinity College Library may lend or copy this dissertation upon request.

Misbah Rizaee

University of Dublin, Trinity College

August 19, 2022

Acknowledgment

I would like to express my appreciation to my supervisor, Professor Timothy Savage, for his constant support and guidance. The success of this research project would not have been possible without his outstanding advices.

I particularly appreciate how approachable and understanding Dr. Melanie Borouche, my strand leader, has always been. Without her assistance, I wouldn't have been able to complete my master's degree.

I am truly thankful to Trinity College Dublin and the School of Computer Science and Statistics for providing me with the opportunity to learn valuable skills and gain experience.

Last but not least, I would like to sincerely thank my loving, understanding, and supportive family for their support. I will always be most appreciative of them. I wish to thank everyone involved.

Misbah Rizaee

University of Dublin, Trinity College

August 19, 2022

Abstract

As the 2030 deadline for sustainable development goals, also known as SDG that were first set by United Nation (UN) in 2015, is getting closer, it is becoming necessary to evaluate how well SDG 10 [1] which ensures equal opportunity and reduces inequalities within and among countries, is being implemented. It is critical that SDG 10 receives special consideration because according to newly released data from the UN at the end of 2021 [2], there are approximately 89 million people who have been forcibly displaced worldwide and nearly 27 million refugees are among them. The UN and Governments are challenged by the growing refugee population as the crisis has expanded geographically and numerically [3].

The reactions to refugee flow often deal with practical challenges [4], such as meeting the basic needs of refugees and organizing the logistics of relocation, but every refugee crisis has its own unique considerations and challenges. Identifying the refugee needs in each refugee crisis has becoming increasingly important and therefore it is a key problem to look into. Equally important is the fact that most existing solutions are post crisis analysis and evaluation which means that the majority of current analysis occurs after the event.

This study investigated the potential of topic modelling on social media to support the identification of refugee needs in near real-time. The research shows that social media and topic modelling have potential to address the current issue in identifying the refugee needs in each refugee crisis as it is occurring. In the past several years, social media sites have played an important role in the area of refugees [5]. In order to overcome the problem outlined earlier, a significant amount of Twitter data was collected through Twitter's API and pre-processed using natural language processing library. Latent Dirichlet Allocation (LDA) which is a statistical model to identify the latent themes in a collection of data was used to classify and identify various topics from the tweets. Following the extraction of the topics using LDA topic modelling, it became clear which fundamental topics were being discussed in the collection of tweets. It revealed the crucial information that could help in providing governments and humanitarian organizations with the ability to quickly meet the needs of refugees. It has been concluded that the data pre-processing approaches helped in increasing the accuracy of LDA model and getting the outcomes that are accurate and relevant to the research question.

The key limitation that must be addressed is that even for a small number of topics, LDA topic modelling takes a long time to estimate. As a result, real-time topic modelling is not possible.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background and Motivation

There is no doubt about the fact that the world is currently experiencing a refugee crisis. The worldwide refugee crisis is unlikely to be resolved anytime soon and it is capturing global attention these years.

There are about 89 million forcibly displaced people and over 27 million refugees [2] who fled their home due to war. Half of them are children [6] and this is clearly having a very traumatizing effect on them, not only because they have witnessed so much violence, they have also been separated from families.

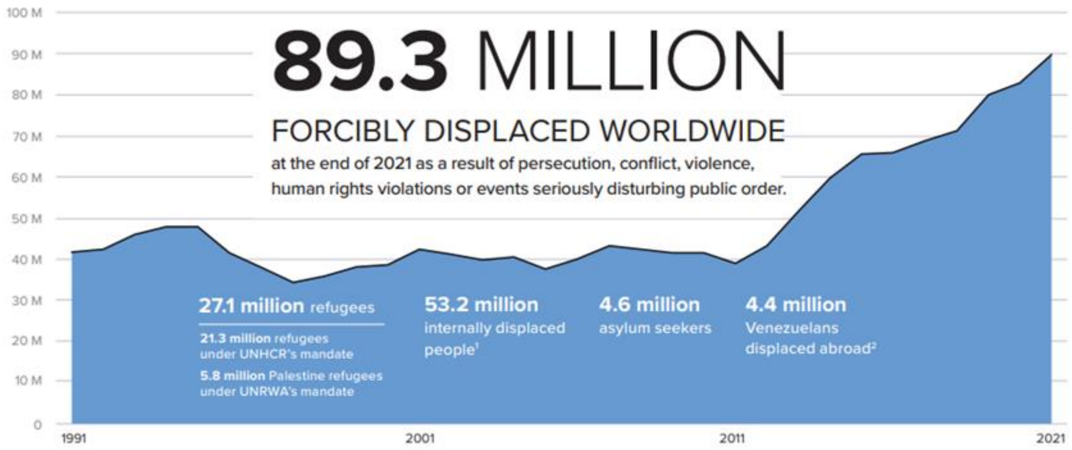


Figure 1. The number of people forcibly displaced is at an all-time high. Image: UNHCR [3]

Like everyone else in the world, refugees require food, housing, and the chance to live. General impression is that the European countries have been incredibly welcoming, they have taken millions of refugees into their homes since 2015 when the Syrian refugee crisis started [7]. The governments and humanitarian organisations have raised funds, mobilized goods and services to support refugees [8].

While the response has been extremely positive, dealing with refugee crisis has not been easy and even few years into the Syrian refugee crisis, the needs have definitely changed and grown over time.

As an example of changing needs over time, Rebecca Hémono et al. [9] completed research from 2015 to 2017 to investigate healthcare providers' opinions on providing health care to Syrian refugees during the humanitarian crisis in Greece. The result of this research shows that healthcare professionals in refugee camps noted a change from physical health issue to mental health issue. This turns out to be even more problematic because the average period of displacement is much longer which means many more needs of refugees change over time or they are not even identified.

Future needs of refugees also may vary, for example there might be even a greater need for smart phones which wouldn’t have been a need 20 years ago. With regard to how the needs of refugees change over time, this problem has attracted attentions and as a result, identifying what the actual needs are in each refugee crisis as it is happening has become an increasing concern and hence a crucial matter to investigate.

When a large volume of unstructured text data is gathered from a social media platform, the challenges in evaluating this data are expected to exist. Manually analysing and categorizing this data takes time and it can be inaccurate. Artificial intelligence and advanced analytics can be used to analyse not only the contents of the tweets but also the sentiment and the emotions behind those tweets [10].

Despite several reviews in the literature that address the importance of social media in assisting refugees and governments, none of the recently published journals have comprehensively discussed the important roles of topic modelling on social media to identify the needs of refugees in near real-time. Only a few studies in literature demonstrate how topic modelling is used to identify the main topics in the social media posts. To fill this gap in literature, this thesis will study the whole area of refugees, social media, topic modelling and sentiment analysis but the angle that is taken on it is use of social media and topic modelling to identify the immediate and emergent needs of refugees in near real-time.

1.2 Problem Description

Having established that the refugee needs will change over time there is now problem of how to identify them. There is a range of analysis done and most of the research in this field is aimed at finding the role of social media in assisting refugees. Social media analysis has been proved to be effective in supporting refugees however they do not use topic modelling to identify the needs of the refugees and this remains an open problem in the area.

Another major problem that requires attention and needs to be addressed is the fact that most of the existing social media analysis is after the event and they have little use during the early stages of a refugee crisis. It was observed recently in the Ukrainian refugee crisis that a large movement of refugees occurred in neighbouring countries such as Poland, Hungary, Romania, the Republic of Moldova, Slovakia and Bulgaria [11]. Many of these countries experienced an unexpected situation which required immediate actions. To develop an effective solution, social media and topic modelling are going to be used to identify the refugee needs in near real-time.

1.3 Research Objectives/Questions

After defining the problems, this section explains the research questions as well as the hypotheses that this project is designed to address. The research aims at finding a solution for this challenging problem of identifying the refugee needs. To address this problem, the following main research question has been formulated:

Research Question: Can real-time LDA topic modelling of relevant tweets identify the current and changing needs of refugee groups?

Aim: To discover the refugee needs accurately in near real-time.

Objective: To use LDA topic modelling algorithm on the Twitter data about refugees.

Hypothesis: The findings of this research are hypothesised to be representative of refugee needs.

Thousands of tweets about the refugees are collected and examined to find topics and themes in response to the desire to understand the refugee needs in near real-time. Topic modelling approach is used to manage such complex data. With the use of this method, it will be investigated to see how topic modelling is used to identify the refugee needs in the topics and which number of topics works best when performance between them is being compared. To extract important insights from the data, its random appearance must be standardised. There are essential processes that must be followed in order to remove noise from the data and create meaningful various topics. Therefore, one of the objectives is to improve the quality of the data.

Chapter 2 Literature Review

There is a refugee crisis going on throughout the world, including, at the time of writing, the Ukrainian refugee crisis [12]. Similar crises have occurred in recent years such as the Syrian refugee crisis in 2015. At the same time, social media has been widely used for integration, sentiment analysis and so on, not only by the refugees [13] [14] but also the countries that they end up in [15].

This section starts by reviewing previous studies on the needs of refugees. The forced immigrations have variety of immediate needs including shelter, education, access to health and etc. It also reviews how social media is used by refugees and governments and how it represents opportunities for them.

It is then followed by exploring various technologies that could be implemented within social media to support the identification of refugee needs. It describes briefly the significance of topic modelling in raising awareness. It also provides a brief overview of studies on topic modelling demand in politics researches. There are also other forms of computer supported analysis conduct on social media, the main one being sentiment analysis. Sentiment analysis aims to help by finding the direction of discussion on social media, change of public opinions and the possibility of further tension.

2.1 Refugee Needs

This section reviews the literature related to the needs of refugees. The previous studies reveal that the needs of refugees are normally shelter, education and access to healthcare. However, each crisis is unique therefore a number of questions regarding what the current needs of refugees are and how to identify them still remain to be addressed.

2.1.1 Shelter

A number of authors have recognized the immediate needs of refugees, some focusing on the housing problems [16]. Housing is an essential need for safety regardless of what type of housing [17] because if housing needs of refugees is not met, the integration of refugees will not be successful. If refugees' safety is constantly threatened, they will be unable to go to the next stage of integration into society [18].

Tents accommodate many of the refugees living in camps today. The normal tent, on the other hand, is unhealthy, unsafe, and difficult to live in for long periods of time [19]. They usually last six to twelve months, but the average period of displacement is more than 15 years [20].

2.1.2 Education

Over half of the refugees in the world are under the age of 18 [21] which means even more children than ever before are spending their whole childhood without the right to an education.

Previous research by Kathleen Fincham [22] has almost entirely focused on the difficulties that refugee children encounter in accessing education, as well as a more basic issue that refugee students experience when it comes to learning [23]. Because of financial difficulties, refugee children are likely to work rather than attend school [24]. It is challenging for refugee parents who are trying hard to afford their families' basic needs to send their children to school. Many refugee children end up working instead of attending school.

2.1.3 Access to Health

Due to language problems and a limited access to emergency healthcare, newly arrived refugees are among the most vulnerable people during a disaster in the world [25]. Research by Antonis A. Kousoulis et al. [26] has shown that Greek healthcare system has for many years been costly but it is much more costly for societies to leave any category of migrant without adequate health coverage. There are 2 reasons for these, the first reason is that migrant who cannot afford professional health services or choose to self-diagnose or self-medicate have a negative ripple effect that endangers individual and public health. The second reason is that refugees who do not have access to low-cost primary healthcare for conditions that could have been prevented or effectively managed at an early stage, face very costly emergency care services later.

3 main areas of refugee needs have been identified however this is always looking at what the needs have been in the past. As every refugee crisis is unique, the needs for the current refugee crisis must be understood which is likely to be slightly different from the crisis in the past.

2.2 Refugees and Social Media

The importance of social media in assisting refugees and governments is considerable. For those who have been displaced, moving to Europe need very detailed information regarding the route and the final destination [27]. Each need raises issues related to where to look for information, whether to believe it or not, and the associated costs. This section of the literature review concentrates on the importance of information seeking and the challenges associated with doing so, how refugees recognize false and misleading information, and the important role of smartphones because they are generating a significant amount data that could be analysed. On the other hand, Information seeking and sentiment analysis have both been extensively used by governments in the field of dealing with refugees and controlling migration by influencing the destination choice of refugees.

2.2.1 Refugees and Social Media

Smartphones and mobile access are increasingly important tools for the refugees who are making their way to Europe [28]. On top of that, social media has a role in more than only spreading information about refugees, it also presents opportunities for them. The findings of a research by Rianne Dekker et al. [29] indicate that refugees' main motivations for using social media in their lives are communication, decision making and access to information.

Refugees have several questions about how they can start a new life in the host countries but they are unsure who to ask for proper answers. According to the findings of the research by Rianne Dekker et al. [29], Syrian refugees prefer social media information that is based on the personal experiences of other refugees. This type of information is often regarded as more accurate information by them.

2.2.2 Governments and Social Media

Social media is considered as a support in migration processes, and it has recently been used as a resource by governments and volunteer groups to encourage integration of refugees and help them in the host countries [30]. An example of social media as a support can be seen in the refugee crisis in 2015 when the European governments and humanitarian organizations started using social media to address concerns generated by forced refugees. There has been research carried out to investigate how social media has been used by different governments. A recent study by Maria Gintova [31] in 2019 concluded that Canadian government use Facebook and Twitter to provide customer service for refugees who are looking to hear directly from government bodies and are expecting specific responses.

On the other hand, social media has also been used as a tool to prevent migration. Jan-Paul Brekke et al. wrote a paper reviewing the use of social media by Norwegian government [32]. He has provided evidence that in 2015 and with the flow of migrants and refugees to European countries, the Norwegian government used social media such as Facebook to influence the destination choice of refugees and prevent migration [33].

Therefore, social media has been powerful for identifying characteristics of each crisis and it has been used for variety of needs in variety of ways however most existing solutions are after the event and therefore have limited use in the immediate time period of a refugee crisis as it is emerging.

2.3 Analysing refugees and social media

2.3.1 Topic Modelling

Topic modelling is a process of pattern recognition in a collection of data. It enables the organization, combination, and visualization of latent topics and patterns found in any type of text corpus. It provides helpful information for future study and helps in improved comprehension of the text [34].

It has been previously recognised that the refugee needs are unique and that most existing social media analysis takes place after the event. The problems that just have been identified in this paper can be solved by using topic modelling as a discovery tool to identify the needs of refugees in near real-time.

2.3.1.1 Raising Awareness Using Topic Modelling

The content of news has a great impact in influencing discourse about issues like the refugee crisis, but the social media sites like Twitter allow stories to be maintained, questioned, and filled with new thoughts and opinions. A study by Adina Nerghes et al. [35] has compared news stories about refugee crisis with the topic analysis of the tweets. The author has shown how Twitter and news converge and diverge. Unlike the news, Twitter has provided a solution by raising awareness and encouraging solidarity and empathy for people in need because the issue of refugees cannot be solved by a single government, ministry or individual.

2.3.1.2 Demand for Topic Modelling In Politics Research

Those who research politics have a need for identifying the topics and there has been research that has looked at analysing large amount of text. Philip Grant et al. [36] proposed research that examines the history of global refugee policy using typewritten and digitally generated documents which are 55,000 pages from worldwide and national archives. The information dates back to the 1970s and has been kept in archives run by the governments of the United Kingdom, United States and the United Nations High Commissioner for Refugees (UNHCR). The overall subject was to examine the roles of the United Kingdom, the United States, and the UNHCR in various refugee situations that happened throughout the 1970s.

2.3.2 Sentiment Analysis

Sentiment analysis is another method that is commonly applied on social media data, but it is not as important as topic modelling, which was previously covered in detail. While sentiment analysis may not have much help in identifying the needs of refugees but it is closely aligned to the area of topic modelling [37].

It becomes easier to make sense of the words and context in online text in order to understand how people feel about a topic or in this case the refugees. One way that sentiment analysis can be performed is to extract known good and bad words from a text and then score the overall sentiment of that text depending on how often good or bad words appear [38].

If some tweets can provide motivation, encouragement and support which can be seen in many tweets regarding Ukrainian refugees, they can be considered as positive sentiment, whereas fear and anger feeling which is usually towards the war, can be categorised as negative sentiment.

2.3.2.1 Direction of Discussion about Refugee Crisis on Social Media

Sentiment plays an incredibly important part when it comes to analysing significant amount of unstructured data that is generated daily on social media. Research by Nazan ÖZTÜRK et al. [39] is about automatically analysing the text in thousands of tweets on Syrian refugee crisis. This analysis was important because it was regarded as a source of information on how people in turkey react and talk about the refugee crisis in their country.

2.3.2.2 Change of Public Opinions on Immigration and Refugees after an Event

Another goal of sentiment analysis is to understand the sentiment on social media posts over time more specifically when an external event occurs. A journal planned to track the sentiment over time which Elizaveta Kopacheva et al. [40] developed, hoping to see change in sentiment as concerns increased when over one million refugees reached Europe in 2015. According to the findings of the research, the refugee crisis in 2015 had a minimal impact on the sentiments of the tweets on this social media site. However, the author has seen a small shift in the negative opinions of the users following the refugee crisis.

2.4 Conclusion

Much study and discussion has been undertaken on what the immediate needs of refugees are at the refugee camps and the countries hosting them but the majority of the academic works which have been reviewed in this section focused on Syrian refugee crisis in 2015.

Although this review has highlighted the importance of social media in identifying things but no study has yet looked at how social media can be used to identify the needs of refugees.

Previous studies have also indicated that LDA topic modelling worked and performed well in the field of dealing with refugees. These studies have demonstrated the potential for identifying the needs of refugees using topic modelling on social media.

As a result, this study is going to create technology for identifying the unique needs within the crisis in near real-time in order to assist the response.

Chapter 3 Methodology

3.1 Design

The goal of this chapter is to provide a high-level overview of the implementation process of the project. Based on the literature, this project has been built which consists of four steps, all of which are described in detail in this chapter. Data gathering using Twitter API was the initial step of the implementation process. Dealing with unstructured data and preparing it for topic modelling required data pre-processing which includes approaches like punctuation removal, tokenization, stop-words removal and token normalization. In addition, it was ensured that the data that has been collected is protected with the proper security measures in place. Following that, various topics from the tweets are classified and identified using LDA topic modelling.

The design of the planned project is shown in detail in the figure below, along with all of the mentioned parts that are crucial for data analysis.

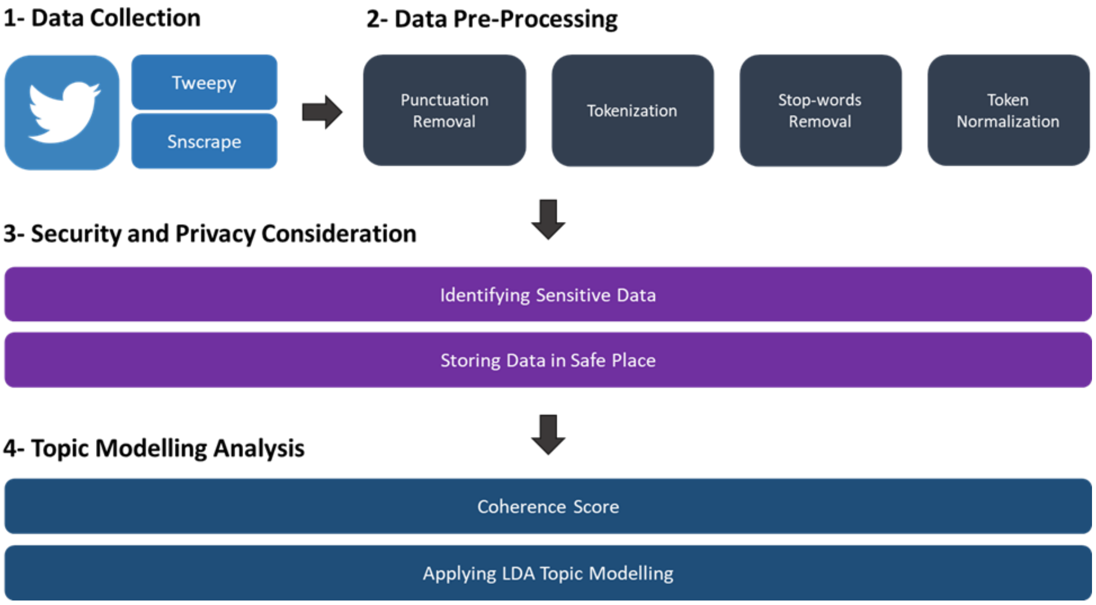

1- Data Collection
Tweepy
Snscrape
2- Data Pre-Processing
Punctuation Removal
Tokenization
Stop-words Removal
Token Normalization
3- Security and Privacy Consideration
Identifying Sensitive Data
Storing Data in Safe Place
4- Topic Modelling Analysis
Coherence Score
Applying LDA Topic Modelling

Figure 2. Architecture design of the project

3.2 Gathering and Pre-processing of Data

In this section of the design chapter, the concentration is on the process of data gathering and pre-processing of the data in order to remove noise from the tweets and increase the general level of data quality. It will be discussed why specifically Twitter has been chosen and the Twitter API libraries that this platform offers.

3.2.1 Choice Of Social Media For Implementation

Twitter is a good choice for data analysis compared to other social media platforms as it is a real-time multi-blogging platform where news appears first and plays an important part in its popularity [41]. People express their opinions and feeling in the posts and comments everywhere on the internet these days [42], and Twitter is the one that always offers the first indication.

3.2.2 Data Gathering

Creating a dataset was the first step in starting the implementation of the project. Selecting an appropriate method for collecting data was very important because there are some limitations that would slow down or stop the process of data gathering. These limitations will soon be covered in more detail. This part outlines data collection methods that have been used in the project.

Twitter offers API platform, to make it easier for the researchers to access the publicly available tweets [43].

Tweepy library: Tweepy is the primary library of the Twitter API [44]. Access to the widely accessible stream of tweets on Twitter is made possible by tweepy. Using this library, new tweets can be collected as soon as they are published. Tweepy has the potential to be an effective tool for monitoring public sentiment on widely discussed topics on social media in real-time.

However, it is tweepy's biggest drawback in this project. Tweepy was basically just designed to collect the most recent tweets from up to 7 days ago [45] but, in this project, the tweets from 2021 need to be collected as well. In addition to the main limitation in tweepy, this library only collects a limited number of tweets in a timeline which means that when the limit is reached, there will a pause before new tweets are once more collected [46].

Snscrape library: Collecting old tweets from Twitter was complicated therefore there has been many studies done on different possible methods for collecting old tweets from Twitter using particular keywords, at a specified timeframe [47]. A useful python library called as Snscrape was discovered in order to collect old Twitter data however when this library was being integrated in the project, there was confusion surrounding it because of the lack of documentation and the problems regarding its development version.

Eventually this has become the default library to collect tweets in the project however, by combining Tweepy and Snscrape, it was possible to get around the API restrictions and collect all the tweets that were needed for topic modelling and sentiment analysis.

It's interesting to note that Snscrape can collect data from multiple social media platforms, including Facebook, Instagram, and Reddit, in addition to Twitter [48]. This is useful if the project is extended to get more data from other social media sites.

Datasets: After deciding on the best library of Twitter API, the process of collecting thousands of tweets started. These tweets are about:

- Afghanistan Refugee crisis in 2021
- Ukrainian Refugee crisis in 2022

These are widely discussed topics in relation to refugees on social media platforms right now. There was additional task involved in this phase of the implementation process, including:

Writing a query using which, the tweets with multiple keywords could be collected. This means that the text in each tweet should contain the words that have been specified in the query depending on how the query is written. For example, to search for tweets that contain either the word "refugee" or "migrant", the string "refugee OR migrant" should have been passed in to the query. Similarly, the string "Ukrainian OR Afghan" collects tweets which contain at least one of these words. Now if these keywords are put in parentheses and a space is added between the two parentheses, the tweets that contain at least one word from each parenthesis will be collected. The following table shows an example of query and the words that the tweets will contain:

Table 1. Snscrape API query

The tweets that will be collected using the query below, contain the following words:						
Query	(Ukrainian OR Afghan) (refugee OR (asylum seeker) OR migrant OR migration OR Immigrant OR immigration)					
Words in the tweets	Ukrainian, refugee	Ukrainian, asylum seeker	Ukrainian, migrant	Ukrainian, migration	Ukrainian, Immigrant	Ukrainian, immigration
	Afghan, refugee	Afghan, asylum seeker	Afghan, migrant	Afghan, migration	Afghan, Immigrant	Afghan, immigration

It should be noted that this query is for the Snscrape API and it is not supported in tweepy. To search for tweets with a specific topic using tweepy, a single word should be passed in to the query.

3.2.3 Data Pre-processing

Every day, Twitter receives millions of tweets [49] and these tweets create a vast volume of unstructured data. One of the most fundamental and basic step of natural language processing is cleaning the data. It will typically increase the accuracy of LDA models by cleaning, summarizing, simplifying, or categorizing text. To remove noise from the tweets, some common pre-processing steps were used, such as:

Punctuation removal: In the first step, the punctuations and any irrelevant information like emoji, special characters and extra blank lines or spaces were removed from the tweets. This is important to remove punctuation before going to the next step of data pre-processing which is tokenization. Let's take the following tokens as an example and assume that the punctuations have not been removed and the following words are left after tokenization step, The issue is that although the last token "it?" which has a question mark at the end of it and the other token "it" which does not have any question mark have the same meaning, but if they are compared together, it is recognised that they are separate tokens therefore they are considered as different tokens. These two tokens will be need to be combined because they have basically the same meaning and later on the effect that they have on the output of the LDA model might not be acceptable. additionally for the word refugees which is followed by a comma at the end of the token "refugees," is the same token as simply "refugees" without comma.

Tokens:

This is a tweet about refugees, isn’t it?

Tokenization: Tokenization is the process of breaking up text on white spaces into useful chunks, and the resulting chunk is known as a token. Tokens act as helpful key components for additional language comprehension. A token can be as small as individual characters or as large as the entire text of document. The most common types of tokens are characters, words, sentences and documents.

Stop-words removal: Many words like “the” and “of” are not very interesting and they are also frequently occurring in the tweets therefore they needed to be removed from the corpus.

Token normalization: The next step is to have one unique word form for different words that are similar or have the same meaning, like some of the words that are currently in use include “playing”, “played” and “plays”. These words basically mean the same thing and they should be merged into one unique form “play” because it does not matter what ending that word has. This process of normalization is known as stemming or lemmatization. Stemming is the process of removing and replacing suffixes in order to reach the original form of the word, also known as the stem. Another story is the lemmatization which the process of analysing the words correctly with the use of vocabularies. It groups the different forms of a word together and returns the base or dictionary form of all of these words which is known as the lemma.

3.3 Data Security and Privacy Considerations

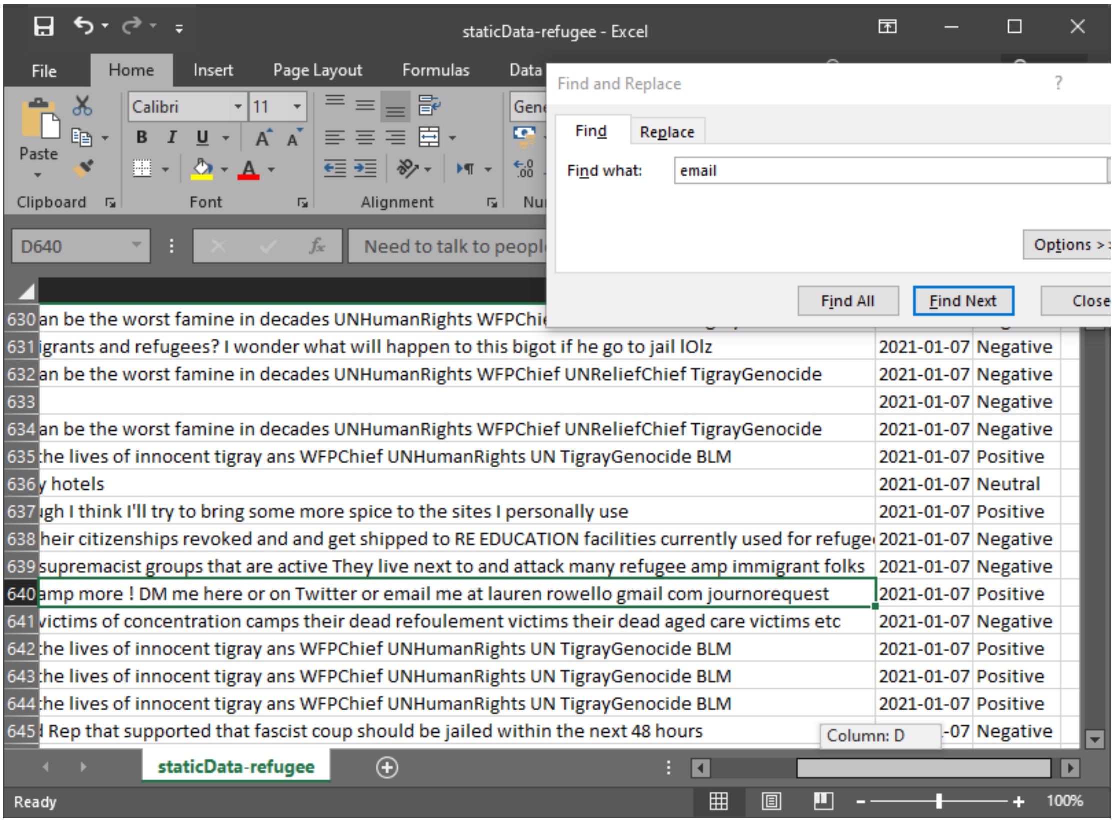
As it was previously mentioned, this research project will look at significant Twitter data and use them for discovering various topics and decision mining in the context of the refugee crisis. Although the data which have been collected from Twitter are already available on the internet and they are open to the public, but they provide certain information about the users. As a result, there are a few key aspects to consider.

3.3.1 Twitter's Safety and Security Features

There are a variety of options on Twitter that allow users to manage who may see their tweets [50]. The easiest approach to keep their tweets safe is to make their Twitter accounts private. Protecting the tweets means that anything they write on Twitter will be private, and their tweets will not be visible in the search engines such as Google or will not be accessible using the Twitter API. Twitter does not reveal sensitive information such as users' email addresses, but many people on Twitter have posted some tweets using their email addresses at some point in their life [51] which will be discussed in the next part.

3.3.2 What Personal Information Can Be Found in The Tweets?

As previously stated, Twitter does not reveal the email address or phone number used to register an account, but many Twitter users do post personal information in their tweets or biographies. The figure below is a screenshot from the tweets that have been collected. As it can be seen, when the phrase "email" is searched, some of the users' email addresses and phone numbers can be found in their tweets.



Find what: email

D640 Need to talk to peopl

630	an be the worst famine in decades UNHumanRights WFPChi		
631	igrants and refugees? I wonder what will happen to this bigot if he go to jail lOlz	2021-01-07	Negative
632	an be the worst famine in decades UNHumanRights WFPChief UNReliefChief TigrayGenocide	2021-01-07	Negative
633		2021-01-07	Negative
634	an be the worst famine in decades UNHumanRights WFPChief UNReliefChief TigrayGenocide	2021-01-07	Negative
635	:he lives of innocent tigray ans WFPChief UNHumanRights UN TigrayGenocide BLM	2021-01-07	Positive
636	y hotels	2021-01-07	Neutral
637	ıgh I think I'll try to bring some more spice to the sites I personally use	2021-01-07	Positive
638	heir citizenships revoked and and get shipped to RE EDUCATION facilities currently used for refuge	2021-01-07	Negative
639	supremacist groups that are active They live next to and attack many refugee amp immigrant folks	2021-01-07	Negative
640	amp more ! DM me here or on Twitter or email me at lauren rowello gmail com journorequest	2021-01-07	Positive
641	victims of concentration camps their dead refoulement victims their dead aged care victims etc	2021-01-07	Negative
642	:he lives of innocent tigray ans WFPChief UNHumanRights UN TigrayGenocide BLM	2021-01-07	Positive
643	:he lives of innocent tigray ans WFPChief UNHumanRights UN TigrayGenocide BLM	2021-01-07	Positive
644	:he lives of innocent tigray ans WFPChief UNHumanRights UN TigrayGenocide BLM	2021-01-07	Positive
645	l Rep that supported that fascist coup should be jailed within the next 48 hours	Column: D -07	Negative

Figure 3. An example of sensitive data found in the tweets

The profile name of the users is along with the other information that were collected from the Twitter. Even if the hackers do not find an email address from the tweets or biographies of a user, they may use email finder tools to get the authentic email address based on profile name.

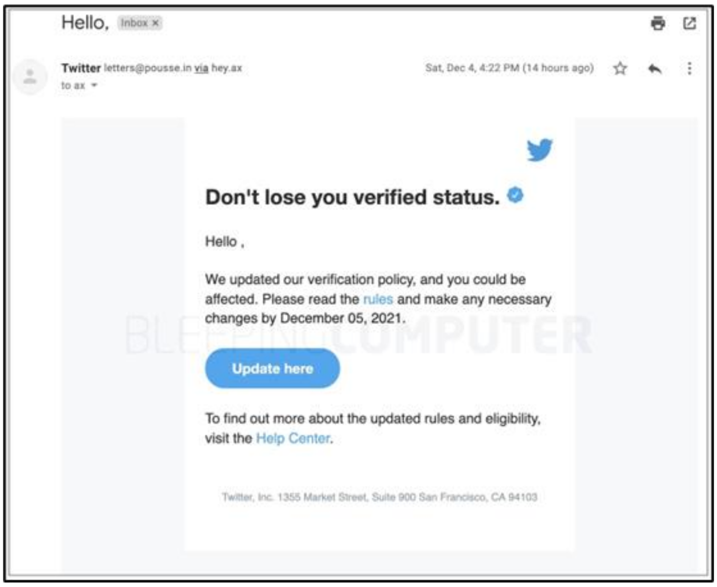
3.3.3 What Happens If the Collected Data Is Leaked?

Personal information is a gold mine for hackers. It's not simply obvious sensitive information like passwords and bank account information that they collect, but any personal information that they can combine to create a profile of potential victims [52]. If the tweets that have been collected get exposed in a data breach, this information then may be used to launch phishing attacks that are specifically targeted

[53]. Phishing emails or direct messages are frequently fake messages that invite receivers to click on a link or respond with sensitive information.

This phishing attack is a typical way for hackers to get access to people's Twitter accounts. Hackers will try to trick their target victim into signing onto fake websites by sending links or messages [54]. The actual purpose is to take personal information, such as email addresses and passwords, in order to obtain access to the victim's account. Since Twitter is such a frequently used social media, hackers often try to send phishing emails posing as Twitter, requesting that users change their passwords [55]. This would allow hackers to steal the user's credentials, which can also be used to log into other accounts.

People have encountered phishing sites that seem very similar to Twitter website in the past. The exact website’s look and visuals is used by hackers to get people to input their credentials. If people reuse the same credentials for many accounts, they may be vulnerable to identity theft.



Hello, Inbox ×

Twitter letters@pousse.in via hey.ax
to ax

Sat, Dec 4, 4:22 PM (14 hours ago)

Don't lose you verified status.

Hello ,

We updated our verification policy, and you could be affected. Please read the rules and make any necessary changes by December 05, 2021.

Update here

To find out more about the updated rules and eligibility, visit the Help Center.

Twitter, Inc. 1355 Market Street, Suite 900 San Francisco, CA 94103

Figure 4. An example of phishing email [82]

3.3.4 How the Security Risks Can Be Reduced?

Although there are many types of phishing attacks, email phishing is the most common and well known [56]. Phishing attacks have extended beyond email and into other communication platforms such as social media and text messaging. Over the last decade, the increased use of social media has made this type of phishing attack more common [57]. The most powerful weapon against all of these attacks is awareness. To keep a secure project, not only the warning signals should be taken into consideration, but also the best practices to follow to restrict the availability of the personal information.

While the collected tweets have offered a significant amount of information for this research project, threats to other areas of the project, such as the user interface (UI) remain high. The initial Idea was to design a UI for this project that would display the result of the analysis as well as the tweets. This idea was changed due to concern about the release of sensitive or confidential information. In this research project, all of the tweets are stored in a Comma-Separated Values (CSV) file and they will not be shared with anyone. The users' identities will be kept anonymous as their names and IDs are not required. The topic modelling algorithm will be applied on the text of the tweets and only the results will be displayed in the UI. So, viewers can only see the data that are authorized to see which in this case it is only the result of the analysis.

Sensitive data management is an approach that consists of data discovery, categorization, regular checking, and protection by combining people, process, and technology. It is a sensible approach to know where the data is stored, what data is at higher risk, who can access it, when these data is accessed and how to secure the data. Most businesses include the following measures in their sensitive data management [58]. Therefore, these strategies have been followed to keep the sensitive data in the project secure.

Table 2. Management of sensitive data

Management of sensitive data in most of the companies	What have been done in the project
Identifying what the company considers to be sensitive data	Similarly, the sensitive data in the tweets has been identified
Knowing where the companies' sensitive information is stored and who can access them	The collected tweets have also been stored in a safe place
Data should be classified according to its sensitivity and the potential for harm to a company if it is stolen	The collected tweets have been classified
Identifying the owner of the data	Twitter users are the owner of the data that have been collected
Identifying if the data is useful or expired, as well as whether it provides an additional concern by keeping them	This data does not have an expiry date and it is necessary to secure them
Data should be deleted as soon as it is no longer needed, or protected if it must be kept	The data is used for educational purposes and will be deleted after the project is finished

3.3.5 Conclusion

All business and developers must do a better way of handling sensitive data. As many hacked companies have discovered, the costs of failing to implement strong sensitive data management procedures are high [59]. If data breaches are caused by poor sensitive data management procedures, then it might take few years to restore the damage, if they can be undone at all [60]. Many people are keeping sensitive data that they do not even realize they have, putting themselves at danger of having it stolen or exposed. Companies should list every piece of information they hold, classify the data, secure it, and restrict access to it. It's terrible enough to get hacked, but losing the personal information of other users in the first place is far worse.

3.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) was first applied by David Blei, a computer scientist in Columbia who was interested to see whether a computer could be trained using a variety of Bayesian learning to detect themes in scientific abstracts from the journal science [61]. It is a statistical model to identify the latent themes in a corpus, a corpus being a group of documents and documents could be anything from newspaper articles to tweets but regardless of the complexity of the documents, LDA simply just views it as a list of words. On its most basic level, LDA takes an input and runs a statistical model and it generates topics and distribution [62].

According to LDA, a quick method to analyse the content of a text is to examine the set of words that it contains. The tweets with similar content will almost certainly use the same set of words. As a result, mining a large corpus of tweets can discover groupings of words that commonly appear together inside the tweets. These groups of words can be simply regarded as trending topics and serve as the foundation for the detailed explanations [63]. Following the introduction, the underlying elements of LDA, such as hyper parameters, generative process and coherence scores are discussed. The focus will be on becoming familiar with how LDA works and learning how to improve the performance of the model.

3.4.1 LDA Plate Notation and Hyper Parameters

Before getting into the specifics of the Latent Dirichlet Allocation model, the terms that represent the model's name will be considered. The term 'Latent' refers to how the model identifies hidden themes in the documents [64]. LDA assumes that the distribution of themes in a document and the distribution of words in topics are Dirichlet distributions. In practice, this leads to better word identification and more accurate topic allocation of documents. The distribution of the themes in a document is referred to as its allocation.

In order to understand how the LDA model works, some of the rules that these models require which are called hyper parameters will be examined [65].

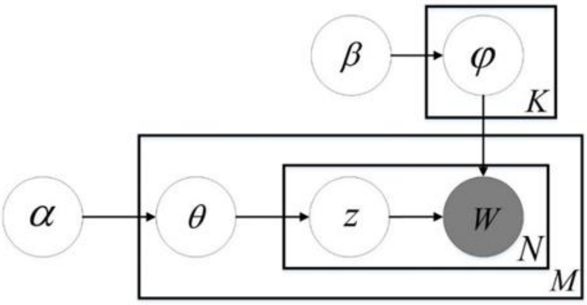


Figure 5. The plate notation representing the LDA model [81]

M and N: The larger rectangle which is denoted with the letter M, shows how many documents there are in the corpus as a whole. The smaller rectangle which is denoted with the letter N indicates the number of words in a document. So, the location of each parameter within these two rectangles determines whether it applies at the document level, the word level, or both of them.

α and β: As it can be seen, the two parameters α and β are present outside of the rectangles and they are called Dirichlet prior. α controls topics in document distributions and β is responsible for words in topic distributions.

A high α value indicates that so many topics are likely to be present in every document, rather than just one or two, while a low α value indicates that fewer topics are likely to make up the majority of a document. Similar to this, a high β value indicates that each topic will contain a combination of the majority of the words but not just any words specifically, while a low β value means that a topic might include a combination of only a few of the words. In other words, high α value will make the documents look more similar to one another, and high β value will make the topics look more similar to one another.

K: The number of topics is usually the most essential hyper parameter, and its selection is determined by the features and the amount of the data in the dataset. For example, the larger the dataset, the greater the number of themes. However, this may not always be the case. If too many documents which are the same to some degree, are added to the original dataset, then that may not include any additional themes. For instance, few thousand tweets on a certain subject.

Selecting a value for K is usually done using estimation. Although this is generally faster to estimate a starting value for k, but it may not be an optimal value. Another approach is to build and train various LDA models with different number of topics k and then calculating the coherence of the LDA models. Visualizing the resulting coherence will help discover the appropriate number of topic k for the model. Finally choosing the value of K with the highest coherence score (this will be discussed in more details shortly).

θ, ϕ: The θ and ϕ are the multinomial distributions. θ is the topic distribution for document M. ϕ is the word distribution for topic k.

Z and W: The letter Z indicates the topic which is assigned to each word therefore making each document. The letter W stands for word.

3.4.2 LDA Formula

Here is the Bayesian network. On the left-hand side, the joint probability of W, Z and θ can be seen. On the right-hand side, there are 5 factors which will be broken down and explained exactly what each component does [66]:

$$p(W,Z,\theta) \ = \prod_{m=1}^{M} p(\theta_m) \prod_{n=1}^{N_m} p(z_{mn} \,|\theta_m) p(w_{mn} \,|z_{mn})$$

Table 3. LDA formula components

Factors	Explanation
$\prod_{m=1}^{M}$	The first component indicates that for each document M
$p(\theta_m)$	Topic probabilities (from the probability p of θ_m) is generated
$\prod_{n=1}^{N_m}$	Then for each word in this document
$p(z_{mn} \,\|\theta_m)$	A topic (with the probability of p of z_{mn} given θ_m) is selected
$p(w_{mn} \,\|z_{mn})$	Finally, a word is selected after a topic is selected. To select a word, the probability of the words in the corresponding topics should be known. This is the probability of the word w_{mn} given z_{mn}. There are a few constraints on this component, first of all it should be non-negative since the probabilities are being modelled and also it should sum up to one.

Table 4. Known and unknown variables

	Status
W - data	Known
ϕ – parameters, the word distribution for topic K.	Unknown
Z – latent variables, topic of each word.	Unknown
$\boldsymbol{\theta}$ – latent variables, the topic distribution for document M.	Unknown

3.4.3 LDA Generative Process

The way that a model like LDA which is a Bayesian model is represented is the same as the way any other Bayesian model is represented through a generative process therefore a generative process for the data is hypothesized and then an inference algorithm has to be derived for doing the inverse problem of learning the actual parameters or a posterior distribution on those parameters of the model that could explain the data.

So, for LDA there is the following generative process [67]. It is assumed that there is K different topics underlying the dataset which means there is K different probability distributions on the same set of words and each of these distributions should capture a theme by putting its probability mass on a coherent subset of the words.

The prior model assumes that each of those topics is generated independently and identically distributed as a Dirichlet distribution. So, this is the prior distribution that is placed on each of the k topics. Each topic is generated once by drawing from this distribution and then fixed for all the time.

$$\boldsymbol{\beta}_k \sim \text{Dirichlet}\ (\boldsymbol{\varphi}),\ \text{for k}\ \boldsymbol{\epsilon}\ \{1\ldots\text{K}\}$$

Then for each individual document, it needs to be decided how it is going to use the topics that are available to it. So, for the mth document, it is done by generating a k dimensional probability distribution, and that is done once for each of the documents in the dataset.

$$\boldsymbol{\theta}_{\mathrm{m}} \sim \text{Dirichlet } (\boldsymbol{\alpha}), \text{ for m } \boldsymbol{\epsilon} \, \{1 \ldots \mathrm{M}\}$$

Now that there is a distribution on the different themes for the mth document, the words that appear in that document have to be generated. For the nth word in the mth document, it is first decided which topic that word is going to come from and this is done by choosing a topic as follow:

$$\mathrm{z}_{\mathrm{mn}} \sim \text{Multinomial}(\boldsymbol{\theta}_{\mathrm{m}})$$

So, z_{mn} will pick out one of the K topics available to it where the probability of picking a particular topic is encoded in $\boldsymbol{\theta}_{\mathrm{m}}$. Once a topic is picked out, the actual word is finally then generated. The nth word in the mth document is chosen by multinomial distributions where it uses the topic picked out by $\boldsymbol{\beta}_{\mathrm{k}}$. Therefore $\boldsymbol{\beta}_{\mathrm{k}}$ is an index or a number between 1 and k that picks out the index of the topic to use to generate the word. This index is going to pick out the correct index for that word in that document.

$$\mathrm{w}_{\mathrm{n}} \sim \text{Multinomial}(\boldsymbol{\beta}_{\mathrm{k}})$$

Table 5. LDA generative process summary [68]

	Process
$\boldsymbol{\beta}_{\mathrm{k}} \sim$ Dirichlet ($\boldsymbol{\varphi}$), for k $\boldsymbol{\epsilon}$ {1…K}	Generate each topic
$\boldsymbol{\theta}_{\mathrm{m}} \sim$ Dirichlet ($\boldsymbol{\alpha}$), for m $\boldsymbol{\epsilon}$ {1…M}	Generate a distribution based on topics for each document
$\mathrm{z}_{\mathrm{mn}} \sim$ Multinomial($\boldsymbol{\theta}_{\mathrm{m}}$)	Pick out a topic
$\mathrm{w}_{\mathrm{n}} \sim$ Multinomial($\boldsymbol{\beta}_{\mathrm{k}}$)	Pick out a word

Now that all of the steps in LDA generative process are covered, few things have been noticed. First, all of the above indicators are unknown except for the data W, Furthermore, it's unclear what the distribution on the topics are as well as what the topics themselves are therefore there are many different things that should be learned with this model.

3.4.4 LDA Model Coherence Score

In order for LDA to work, it is necessary to decide how many topics are going to be discovered [69]. It is important to get a general understanding of how many topics there are even before building an LDA model and applying topic modelling to the data.

It is possible to build several LDA models with different number of topics (k). The topic coherence score indicates how well a topic model generates coherent topics. As a result, the topic model with the highest coherence score can be chosen because better topic model is indicated by a higher coherence score [70]. There are also other additional methods, such as Perplexity, to examine the performance of the LDA models, however they are not as good as coherence [71].

First of all, a fixed number of K was chosen for the LDA model, which was 5 and its coherence score was calculated as a baseline. This gave a coherence score of 0.2746. Then the LDA model is ran again just like it was done before but instead of specifying a single value of K, a range of values from 3 to 49 was specified. This took a while to run each model because depending on the size of the corpus and the speed of computer, a topic model could take few minutes. Once the output of coherence scores was plotted, there were various goodness-of-fit measures that further helped interpreting the output.

3.5 Model Fitting

Once the most optimal value of K that best fits the model was figured out, the latent Dirichlet allocation which is built into Gensim Package was performed. In addition to the number of topics, Dictionary and corpus are the two basic parameters to the LDA model. Each word in a document is assigned a specific id by Gensim. The above-mentioned corpus is a mapping of (word id, word frequency). For example, the output (0, 1) indicates that word with the id 0 only appears once in the first document. There are more different parameters when LDA model is called [72] but the most important one was the number of topics. The other parameters include:

Chunksize: This parameter determines the number of documents analyzed at once by the training algorithm. Increasing this number will increase the speed of training.

Passes: It determines how often the model is trained on the entire corpus. "Epochs" is another term for passes. Iterations is a technical term that refers to the number of times a loop is performed over each document. It's critical to have a sufficient number of "iterations" and "passes".

Alpha: The Dirichlet prior which is used in the model is controlled by this parameter. If alpha is set to a value close to zero, the LDA model will use fewer topics per document but if alpha is set to a larger value, the LDA model will use more topics per document. This alpha parameter is adjusted automatically if it is set to auto.

3.6 Summary

All the techniques and methods that have been used to carry out this research were outlined in this chapter. A summary of each task in data collection phase as well as the pre-processing phase are covered. Two datasets have been studied based on their content and the same analysis is performed on both of them in order to look into the differences between two refugee crises.

Additionally covered were the rules and formulas that LDA models use which gave a better understanding of how these models work. The effectiveness of these models is influenced by the choice of optimal hyperparameters. Since they affect the performance of the LDA model, the important hyperparameters in LDA model and the method of determining the most optimal value of K parameter have been studied.

Chapter 4 Results

This chapter is structured around the methodology chapter. It will concentrate on the findings of each step in the implementation of the project. It is discussed what makes Snscrape the best Twitter API for tweet collection. It is explained how data pre-processing has a great impact on the performance of LDA models and what happens if the data is not pre-processed. For easier comparison, the results of all coherence scores are displayed in figures and tables. Finally, the latent topics in each dataset are presented in two tables. A brief explanation of the tool used to see the outcome of the LDA model is also provided.

4.1 Data Gathering

Accessing Twitter data using tweepy and Snscrape seemed inefficient given that the tweets might get collected twice but the purpose of using two libraries lies in their responsibilities in the project.

Snscrape has separated itself as a library that enables researchers to scrape tweets without the limitations of Tweepy despite Twitter making updates to their API. Therefore, this library was used for scraping as many tweets as it was needed without any limitation.

Another main use of this library is the ability to scrape historical tweets which was found very useful in the project. This feature was used to collect the tweets about Afghanistan refugee crisis which was a widely discussed topic on social media in 2021. This feature added two new arguments in the Snscrape query to specify a timeframe. These arguments include ‘since’ and ‘until’.

On the other hand, tweepy serves a different purpose in the project. It was mainly used to collect live tweets directly from Twitter and apply sentiment analysis on them in real-time. This is one of the most interesting things about this library because everybody on Twitter is continuously discussing things, ideas and anything else in between. Overall, there are many positive outcomes created from these libraries of Twitter API that improved the data gathering process.

25,157 tweets have been collected about Afghanistan refugees and 10,562 tweets have been collected about Ukrainian refugees. The collected tweets were published during a certain period of time. For instance, the tweets regarding Afghanistan refugees that were published only between 1st of January 2021 until 31st of December 2021, have been collected. Similarly, the Ukrainian refugee dataset contains the tweets that were only published between 1st of January 2022 until 14th of March 2022. It is necessary that each dataset contains extensive detail and captures a range of needs that the refugees have before, during and after an event.

4.2 Data Pre-processing

Data pre-processing step can be considered as one of the most important steps of building this project right now. The more quality data and machine learning algorithm or statistical method used in this project, the better the LDA model performed. If the LDA model does not perform well, the experience is that the data is not still ready to be analysed and therefore it needs to be converted into a more understandable data.

It was obvious that the data taken from online social media platforms is raw unprocessed data which means it has false, incompleteness or inaccurateness. It was required to convert this data into a readable format that can be analysed and predicted. The example below is a tweet which is taken from the Ukrainian refugee dataset. The first thing that draws attention is the use of abbreviations and shortened form of words in this tweet. People frequently use the combination of few words that is shortened by removing letters and adding an apostrophe. For instance, “doesn't” is a shortened form of the words “does not”.

Another thing is that people misspell words in the tweets and they can still understand what that tweet is about, but computers do not have the ability to understand misspelled words in the same way human beings can understand. In the example below there is one misspelled word which is “*freind*".

Example:

“*The Ukraine situation is truly terrible Currently a freind of my mother is making an attempt to flee the country in order to seek refugee at our house It's super scary and I really hope this situation doesn't get any worse*”

Emoji detection is not very difficult, but they caused issues in the pre-processing step. The biggest issue faced here was that there were several emojis that were missing after pre-processing the entire dataset. This is because all emojis are created using different platforms, for example some people use a computer desktop and other people use android phone or iPhone to publish a tweet. Each of these platforms supports and manage Unicode in various ways [73]. Depending on the way these platforms render certain emojis, these emojis might not seem exactly the same and therefore they cannot be recognised in the pre-processing step.

The missing emojis were not detected until after a visual inspection of the result. The emojis that were missed in the pre-processing step were found among other words in the topics. This problem was solved by checking more Unicode to find all of the emojis that are created using different platforms.

4.3 Comparison of LDA Models

The use of topic modelling to organise unstructured text data is interesting, but since the number of topics are not always easy to estimate, topic coherence measures have been recommended to help decide between good and bad number of topics.

Topic coherence is a method that can be used to evaluate a topic model that is easier for people to understand. This approach examines a collection of words in created topics and scores how interpretable these topics are.

Several metrics that compute coherence in different ways exist, but Cv shows to be the one that is most compatible with human interpretability. It is a formula that has been shown to correspond well with human decision-making. It counts the number of times the topic words appear together in the corpus. Cv is one of the topic coherence measures supported by Gensim [74].

The measurement of topic coherence scores was found very useful for comparing different topic models based on their accuracy. The decision making became much easier when there was topic coherence because it gave a clear picture of how good LDA models with different number of topic (K) performed.

4.3.1 Comparison of LDA Models on Ukrainian refugee dataset

According to the graph below, which displays the coherence scores for the Ukrainian refugee dataset, when the number of topics was 12, the highest coherence score was achieved which was 0.3824. Based on these considerations, it was determined that 12 was the most optimal number of topics for LDA model.

The coherence score line rises as the number of topics increases, but it is decreasing between 12 and 13. The number of topics that is chosen will still be determined by the needs of the project. Although the coherence score was also high when the number of topics was 45 but this number was large and would result in having so many topics that would contain repetitive words.

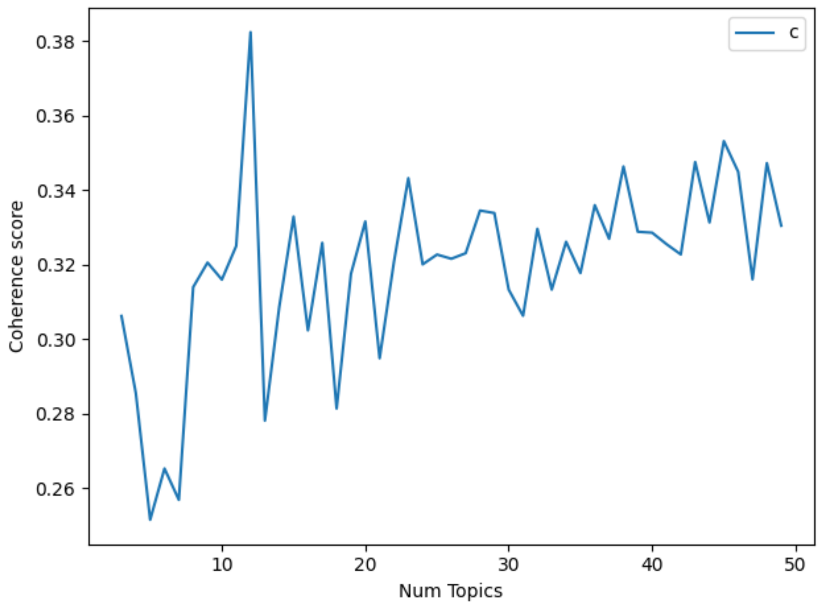


Figure 6. Determining optimal number of topics for Ukrainian refugee dataset

Table 6. Coherence score for each value of K for Ukrainian refugee dataset

Value of K	Coherence	Value of K	Coherence	Value of K	Coherence	Value of K	Coherence
3	0.3062	16	0.3023	29	0.3338	42	0.3227
4	0.2855	17	0.3258	30	0.3133	43	0.3475
5	0.2515	18	0.2814	31	0.3063	44	0.3313
6	0.2653	19	0.3175	32	0.3296	45	0.3531
7	0.2569	20	0.3316	33	0.3133	46	0.3449
8	0.3139	21	0.2949	34	0.3261	47	0.316
9	0.3205	22	0.3208	35	0.3177	48	0.3472
10	0.316	23	0.3432	36	0.3359	49	0.3304
11	0.325	24	0.3201	37	0.3269		
12	0.3824	25	0.3227	38	0.3463		
13	0.2781	26	0.3216	39	0.3288		
14	0.3086	27	0.323	40	0.3286		
15	0.3329	28	0.3345	41	0.3255		

4.3.2 Comparison of LDA Models on Afghanistan refugee dataset

For Afghanistan refugee dataset, 9 was found to be the most effective number of topics to use in order to improve the performance of topic modelling after examining the topic coherence over a range of topics from 3 to 49.

Since having just one or two topics makes no sense and would result in all of the words being on just one or two topics, the range of topics starts at 3. Figure 7 shows that the coherence score line rises as the number of topics increases until it reaches 9.

Generally, it is important to take into account both qualitative and quantitative aspects when deciding the number of topics and analysing the interpretability of the topic model.

This means that once the optimal number of topics has been decided, the next thing to consider is how to accurately evaluate and improve the interpretability of those topics. To check if the outcomes of the topic model make sense for the use case, one method is to visualize the findings. The LDAvis tool can be used to see how the LDA model fits various topics and their top words.

The visualization should include just the right number of topics to allow readers to identify between major themes, but not too many that make the topics difficult to understand. Let's assume that a greater number is found to be the optimal number of topics and it results in having so many topics, in that case, the number of topics is not fixed and the topic model can still improve in terms of interpretability by rerunning the coherence measurements.

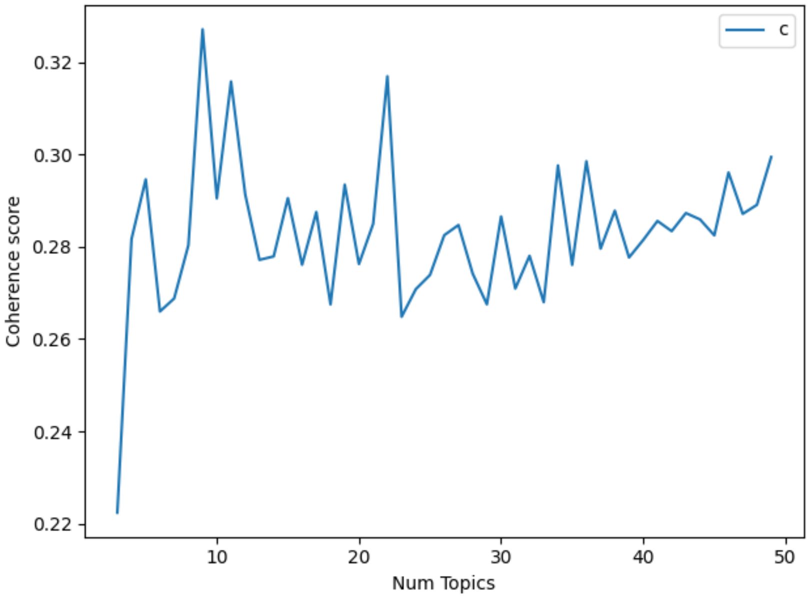


Figure 7. Determining optimal number of topics for Afghanistan refugee dataset

Table 7. Coherence score for each value of K for Afghanistan refugee dataset

Value of K	Coherence	Value of K	Coherence	Value of K	Coherence	Value of K	Coherence
3	0.2224	16	0.2761	29	0.2675	42	0.2834
4	0.2817	17	0.2876	30	0.2866	43	0.2873
5	0.2946	18	0.2675	31	0.271	44	0.2859
6	0.266	19	0.2935	32	0.2781	45	0.2825
7	0.2688	20	0.2763	33	0.268	46	0.2961
8	0.2804	21	0.285	34	0.2977	47	0.2872
9	0.3272	22	0.317	35	0.2761	48	0.2891
10	0.2905	23	0.2648	36	0.2986	49	0.2995
11	0.3159	24	0.2708	37	0.2796		
12	0.2914	25	0.2739	38	0.2879		
13	0.2772	26	0.2825	39	0.2777		
14	0.2779	27	0.2847	40	0.2815		
15	0.2906	28	0.2742	41	0.2856		

4.4 Results of LDA Model

4.4.1 Result of LDA Model on Ukrainian refugee dataset

Topic one for example contains the words spare, room, house and property therefore this topic is about physical house. All the topics nearby are also about housing needs of Ukrainian refugees, for example topic two in particular seems to be about a scheme to accommodate Ukrainian refugees because it contains the words government, home, scheme and host. Similarly, topic 4, 5 and 6 seem to be more or less the same thing. They overlap each other therefore they are semantically related and they focus mostly on healthcare hence the healthcare information has been spread across these three topics.

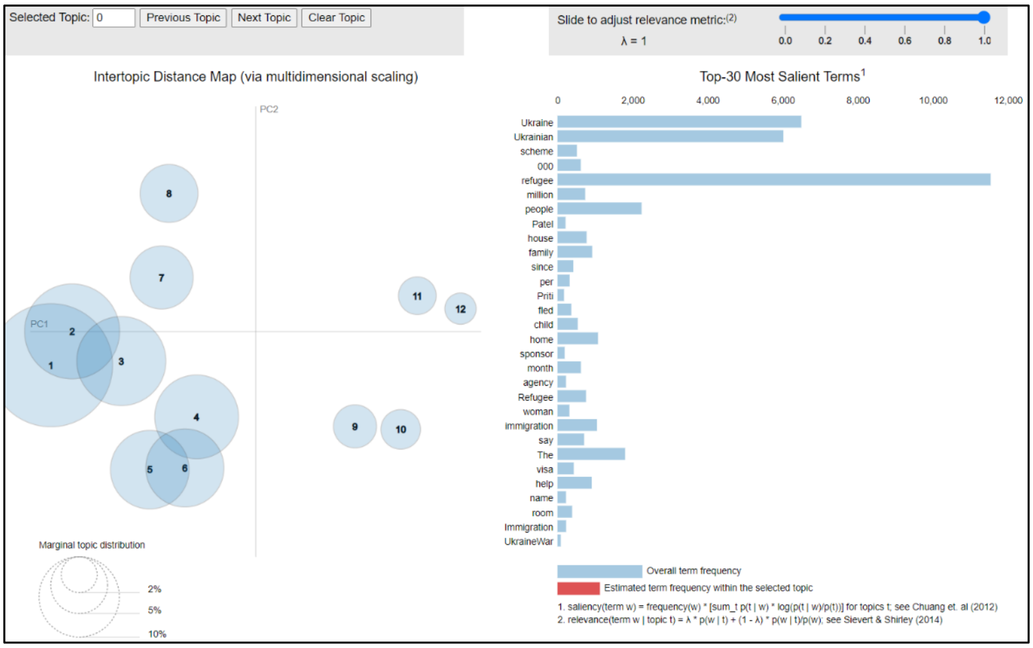
Table 8. Labelled topics of Ukrainian refugee dataset

Topic No.	Label	Words				
		1	2	3	4	5
1	**Physical Housing**	Spare	Room	House	Bedroom	Property
2	**Housing Provision**	Government	British	Home	Scheme	Host
3	**Transportation**	Border	Poland	Pick	Driver	Walking
4	**Healthcare**	Healthcare	Border	Migration	Refugee	War
5	**Healthcare**	NHS	Elderly	Organisation	Government	Refugee
6	**Healthcare**	Unvaccinated	Street	Homeless	War	Invasion
7	**Winter Aid**	Cold	Help	Survive	Struggling	Refugee
8	**Family**	Family	Child	Women	Mother	House
9	-	Volunteer	Donating	Team	Organization	Labour
10	**Education**	Education	Kids	UNHCR	Refugee	Immigration
11	**Food**	Food	Women	Baby	Homelessness	Ukrainian
12	-	Moldova	Math	Art	Church	Refugee

So, several of these topics have been labelled except topic 9 and 12 because the words in these topics are the ones that unless they are shown in combination, it would not be possible to figure out what documents containing these words. So, if a document is about Moldova, then that document could be related to the Ukrainian refugees in Moldova but when the document has a mixture of other words together like math and art, it becomes clear that this is a topic about what Ukrainian children are being taught in the host countries.

It is generally difficult to look at the lists of top words from the LDA model but there are some tools available to support this visualization. LDAvis is an interactive tool and a useful package to visualize the output of LDA model. It was initially a R package however it has since been ported to Python [75]. It gives an HTML output that can be opened in a browser.

In the following visualization, there are two views to look at the topic model, the first one on the left is a map and the idea here is that each topic is represented as a circle and the size of the circle corresponds to how much of the document collection this topic explains. The circles that are closer to one another represent topics that are semantically related therefore it becomes easier to find what topics are in the data and how they are related. The farther apart they are, the better the outcome is. In this visualization 6 out of 12 topics overlap each other but not too much which is a good result.

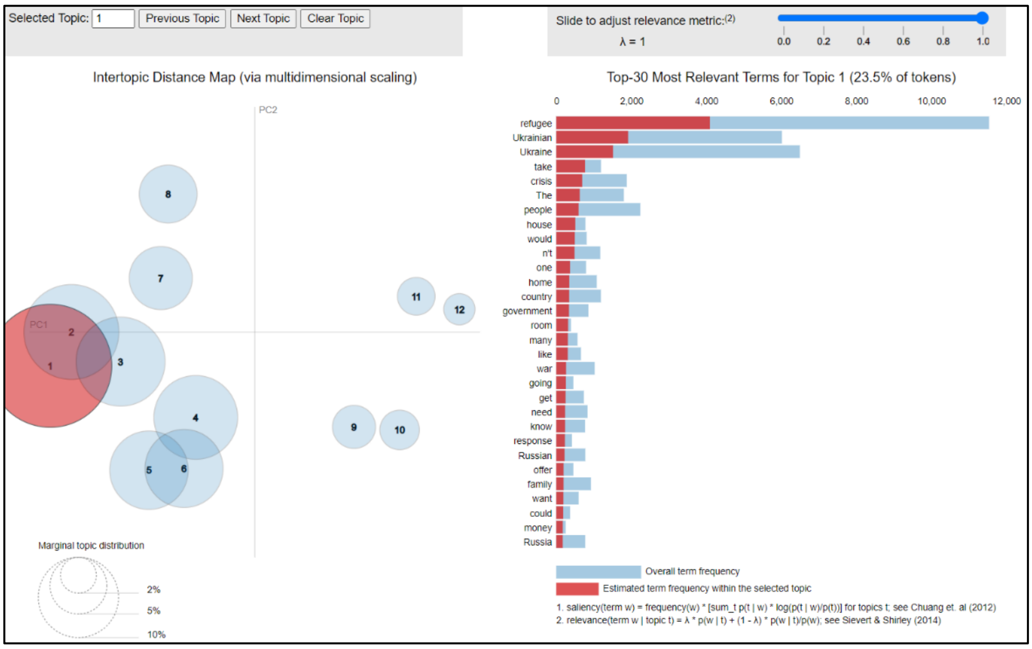

Selected Topic: 0

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. The layout of LDAvis for Ukrainian refugee dataset

When a given circle or topic is chosen, a description of that topic in terms of its top words is shown on the right. The words that are displayed here are ranked in terms of how important they are to this topic.



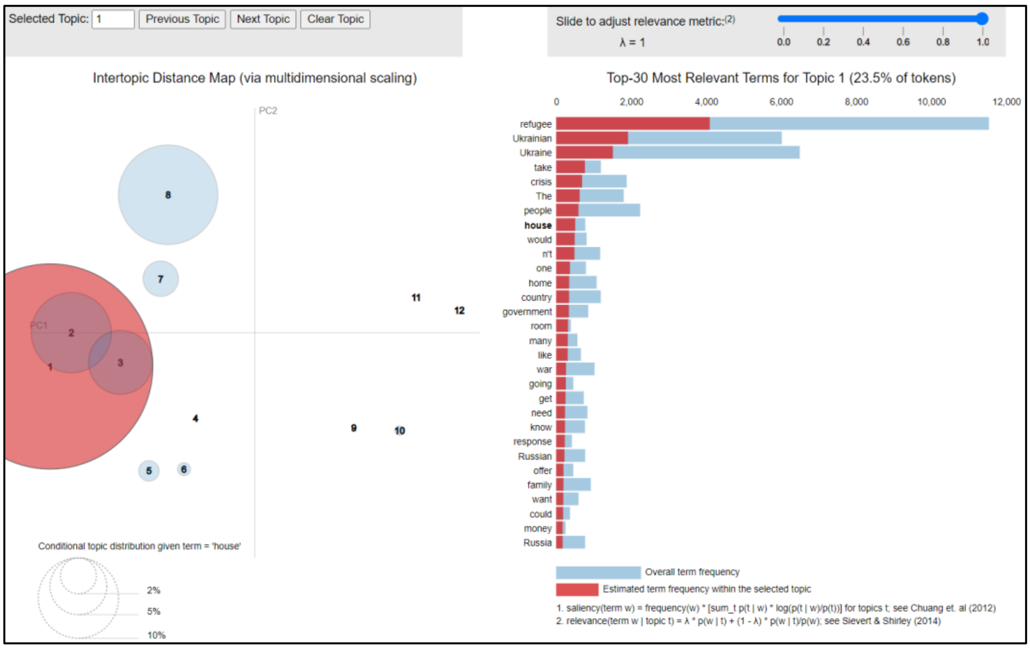
Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9. List of top words in a given circle or topic in LDAvis for Ukrainian refugee dataset

This visualization contains further features for considering which words are the most unique to this particular topic and how each word relates to other topics. When a word is highlighted in one topic, the other topics where this word is frequently used in are also highlighted. As can be seen in the figure below, when the word “house” gets highlighted, the size of circle 1 and 8 get bigger while other circles get smaller which means this word has been used frequently in topic 1 and 8.



Selected Topic: 1

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 10. LDAvis feature to show relationship between word and topics for Ukrainian refugee dataset

4.4.2 Result of LDA Model on Afghanistan refugee dataset

The table below lists 9 topics that were found in Afghanistan refugee dataset, along with 6 representative words from the most important words for each topic. The table also shows that people on Twitter talk about a variety of topics, 8 of which are relevant to the refugee needs while 1 topic on “America” seems irrelevant to the area of this research.

6 out of 8 relevant topics, including “Winter Aid”, “Education”, “Housing”, “Food”, “Healthcare” and “Gender inequality” are clearly related to the refugee needs, while 2 other topics including “Neighbouring Countries” and “Border” could be just considered as topics that are generated from Afghanistan refugee dataset since they are about the countries that receive the most refugees from Afghanistan but they are still closely related to the Afghanistan refugee crisis.

The results of the LDA model are interesting because they show that despite the refugee crisis, people support women in the fight against gender inequality and they pay attention to how important gender justice and women's rights are in this circumstance.

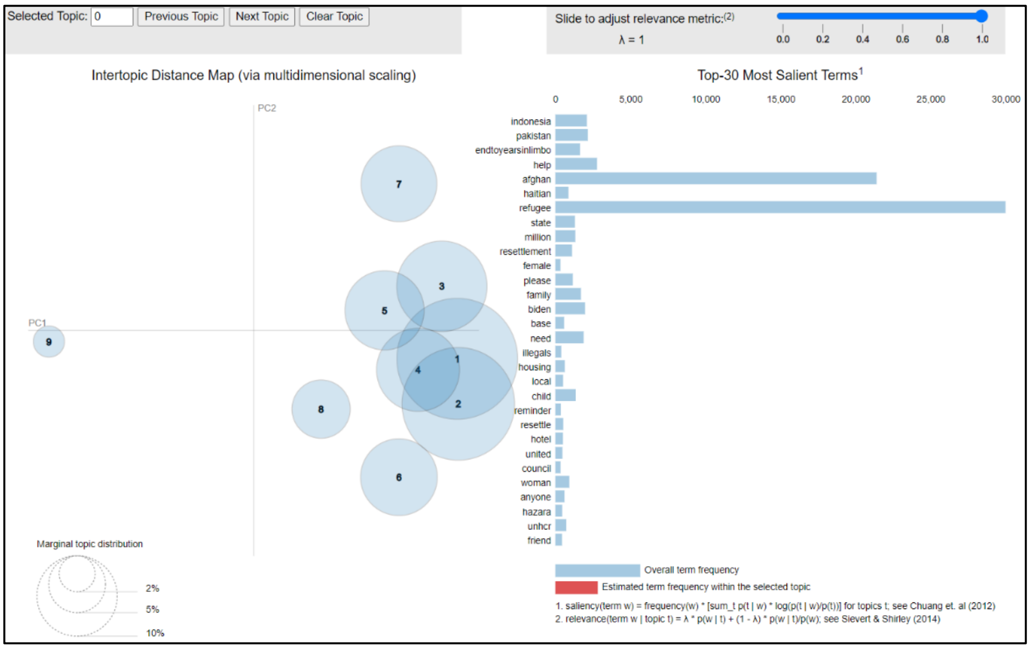
Table 9. Labelled topics of Afghanistan refugee dataset

Topic No.	Label	Words					
		1	2	3	4	5	6
1	**Border**	Border	Camp	Humanitarian	Resettlement	Donate	Refugee
2	**America**	Joe Biden	America	Democrat	Dollar	Refugee	Money
3	**Winter Aid & Religious Minorities**	Clothing	Clothes	Winter	Meal	Hazara	Religion
4	**Education, & Gender Inequality**	College	University	Woman	Family	Vaccinated	Afghanistan
5	**Neighbouring Countries**	Pakistan	Iran	India	Hosting	Refugee	War
6	**Housing**	House	Welcome	Home	Homeless	Room	Resettlement
7	**Food & Job**	Work	Job	Food	Family	Need	Aid
8	**Healthcare & Family**	Woman	Child	Kid	Healthcare	Treatment	Refugee
9	**Gender Inequality**	Female	Girl	Sex	Slave	Refugee	Border

LDAvis package was also used to visualise the topics in the dataset for refugees from Afghanistan. The output again shows an interactive chart that displays 9 topics and their top words in order of importance to each topic.

Much like the earlier LDAvis figures, on the left side, the topics are displayed as circles. A strong LDA topic model will have large circles, non-overlapping circles spread across the graph rather than being clustered in one quarter of the graph.

As it can be seen in the figure below, the size of the circles is often large which is a good result but 3 circles overlapping each other too much. Another negative point to note is that the circles are not spread across the graph and they are mostly in one half of the graph however, the model and its visualization produced a good picture of the key topics in the dataset for refugees from Afghanistan.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 11. The layout of LDAvis for Afghanistan refugee dataset

4.5 Conclusion

A wide range of topics that are discussed in relation to refugee crisis in Ukraine and Afghanistan on Twitter are found using 35,719 collected tweets, pre-processing methods and the LDA model.

The topics that have been found can address the basic needs of refugees, such as housing, food, healthcare and so on. It also addresses the fact that winter is the most difficult and dangerous time of the year for refugees. The cold temperatures in winter could affect the health and housing needs of refugees, increasing danger for those who are already at risk. In terms of gender inequality and the right to access education, people on Twitter have expressed concerns about refugee women and children.

While the LDA model has been recognized for its abilities to accurately and clearly capture words and latent topics in the data, it is important to also address its main drawback. The complexity of the LDA model is mostly dependent on how many topics to include. Given how computationally intensive LDA model is, it takes a long time to compute models with different number of topics and measure their coherence scores.

Chapter 5 Discussion and Conclusion

This chapter is divided into two sections. In the first section, there is a discussion of how the findings of the investigation influence the hypothesis. The important findings of the research are also summarised.

In Conclusion section, it will be discussed how the results correspond with the expectations and how they can be used. The unusual things that were found while carrying out the research and the limitations that appeared will be explained. Following the conclusion of the present research, the future work will be discussed.

5.1 Discussion

The hypothesis posed at the beginning of this document which claimed that the findings of this research would be representative of refugee needs is now supported by the topics that have been extracted since they are relevant to the needs of refugees.

The present findings confirm that not only the needs of refugees change over time but also, they are unique. For instance, in comparison to the Afghanistan refugee dataset, the dataset for Ukrainian refugees shows no evidence of gender inequality and Religious Minorities.

In addition, these findings provide additional information about, Gender inequality language which is often repetitive in Afghanistan refugee dataset, with the same words appearing in many discourses for very different purposes. This conclusion is based on the fact that many tweets in Afghanistan refugee dataset are expressing concern for Afghan women.

Gender equality is also one of the 17 Sustainable Development Goals set by the United Nation [76]. There were conversations regarding how crises like wars can influence women's lives. It is reasonable to say that achieving Sustainable Development Goal 10 which ensures equal opportunity and reduces inequalities within and among countries will help in the achievement of the other Sustainable Development Goals such as Gender equality.

When combined together, these topics could serve as a useful summary of the main needs that the refugees have in any refugee crisis. For instance, winter aid is a common need that refugees from Ukraine and Afghanistan had in both crises.

5.2 Conclusion

The LDA model performed well overall, making it a strong choice to model the topics of a large dataset about refugees, however it is not quite real-time due to topic modelling computation time and it might take about 30 minutes to complete. It has been determined through applying LDA topic modelling on the Twitter data about refugees that the latent topics that have been discovered are representative of refugee needs.

The results can be used in the following ways:

Focus Areas: The findings of this research can be used to highlight which groups of refugees require the most assistance, enabling the United Nation and humanitarian organizations to focus their efforts. For example, a particular focus on the safety of women and girls is suggested as the model highlights that they are frequently confronted with greater levels of danger and they are more frequently victims of gender-based violence.

Emergent Issues: It could be used for providing insights into certain situations. For instance, if this analysis is applied on a weekly basis by a member of a refugee group and it is noticed that there is an increase in some issues related to sickness, it can be considered having an epidemic of some type of disease among refugees in a way that it wouldn't immediately be detected without running this analysis.

Policy Confirmation: In addition, the results could be used for confirming an existing policy. For instance, the United Nation assists refugees in achieving their access to state services, such as public health services, access to education and so on, when it is necessary under its urban refugee policy [77].

5.2.1 Limitations

The followings are meant to demonstrate the limitations that were encountered in this research project. These limitations mainly belong to the data gathering phase of the project. To prevent delays and errors, these limitations must be taken into account while building the project.

Tweepy Rate Limit: There are rate restrictions on how frequently API calls can be used in the Twitter API. To be exact, 900 API calls are permitted every 15 minutes [78] but the restrictions are different based on the account tier, for example the limited number of API calls in this project was 50 with no more than 100 tweets in each call. When going above this limitation, waiting between 5 and 15 minutes was required before calling the API again.

Tweepy Timeline Limit: In addition, Tweepy restricts users to the most recent tweets in a timeline, for example, the tweets that were only published during the previous 7 days. An academic research account is necessary if a user needs to go back further to collect older tweets [79], but Twitter does not easily provide them access to academic research account. As compared to the developer account that was created for this project in order to be able to collect tweets using tweepy, a considerably longer approval process and much higher refusal rate are expected when creating an academic research account.

Disabled Location in Tweets: The ability to track the location of the tweets is limited. Twitter users must choose to manually add their location or enable Global Positioning System (GPS). According to an estimate, only 1% to 3% of Twitter users have enabled accurate location tracking using GPS [80]. Many users follow the security procedures like deactivating geolocation hence, geolocation information is absent from many tweets. This indicates that it could not truly be helpful for collecting tweets based on their location.

Topic Modelling Computation Time and Memory Requirements: LDA topic modelling takes a long time to estimate even for a limited number of topics. It takes about 30 minutes to measure the topic coherence scores and run the LDA model in a standard computer without using Graphics Processing Unit (GPU) or additional memory.

5.2.2 Future Work

The goal of this research was to find the latent topics in the collection of tweets using topic modelling and identify the refugee needs in those topics. Incidentally, sentiment analysis was carried out to see if any correlations can be found. Sentiment analysis is a good area for future research to understand changing perceptions of refugee crisis over time.

By using sentiment analysis, the emotions that were expressed in the tweets were determined. They were either positive, negative or neutral. With sentiment analysis, it was possible to understand and interpret public opinion because the datasets were large and sentiment analysis makes it more efficient as well as consistent to analyse these datasets.

In addition to analysing the sentiment of the datasets, a dashboard has been designed which performs sentiment analysis on live tweets that are collected using Tweepy library. The result of the sentiment analysis on the live tweets is presented in real-time using a Highchart (see Appendix C for dashboard)

The following bar plots make it possible to determine how people generally feel about refugee crisis in Ukraine and Afghanistan. Surprisingly, the dataset on refugees from Afghanistan has slightly more positive tweets than negative ones. This might be as a result of the tweets in the Afghanistan refugee dataset being from a longer time period. They were posted between 1st of January 2021 until 31st of December 2021.

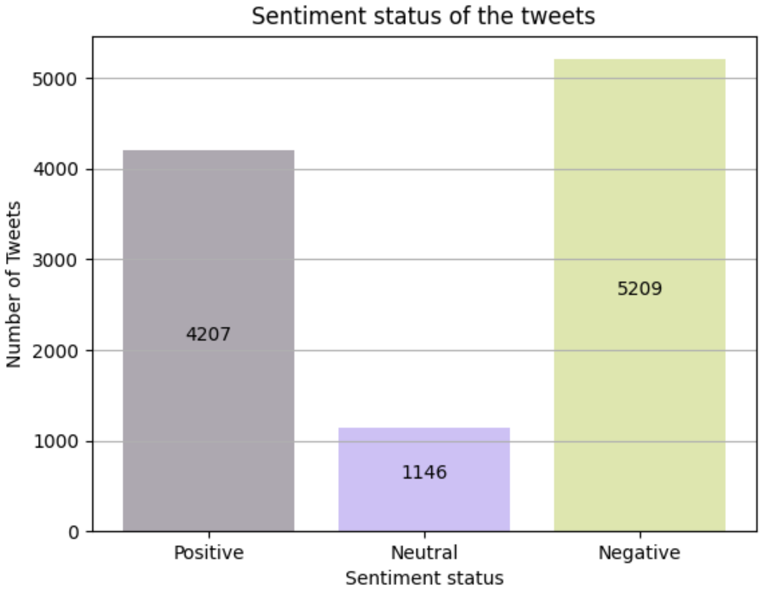


Figure 12. Sentiment status of the tweets in Ukrainian refugee dataset

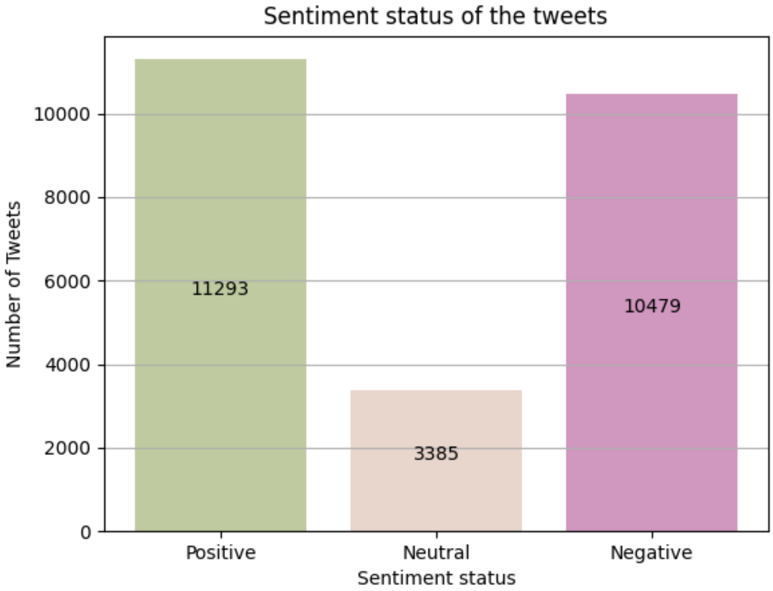


Figure 13. Sentiment status of the tweets in Afghanistan refugee dataset

Sentiment analysis was used in variety of different contexts, for example, the strength of sentiment analysis became obvious when additional data factors such as number of retweets, were taken into account.

The number of times each tweet has been retweeted is included in the Twitter metadata. This is an interesting pattern that was found by examining how the overall number of retweets changed with the sentiment.

In Ukrainian refugee dataset the Twitter users pay far more attention to and retweet more frequently those tweets with a Positive attitude. While in Afghanistan refugee dataset, it appears that negative contents on Twitter significantly increases users' engagement.

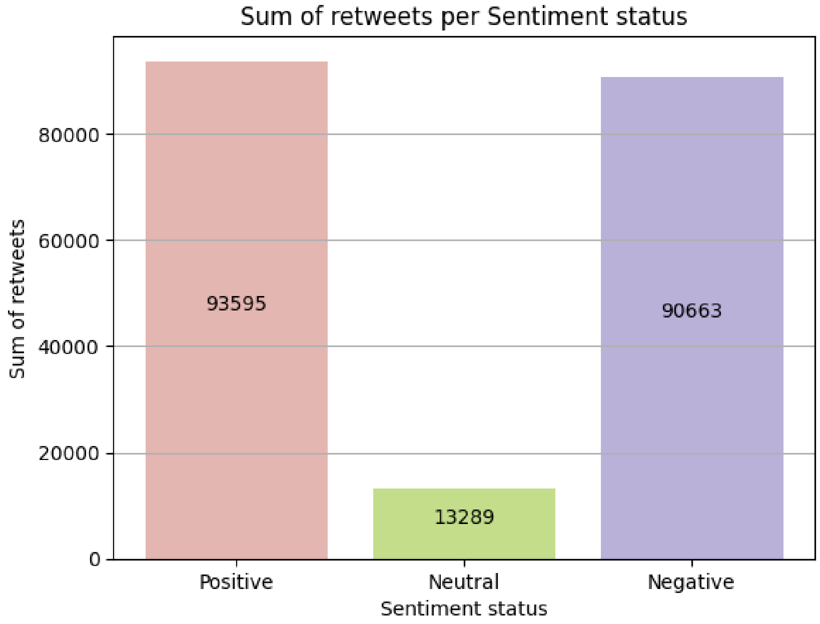


Figure 14. Sum of retweets per sentiment status in Ukrainian refugee dataset

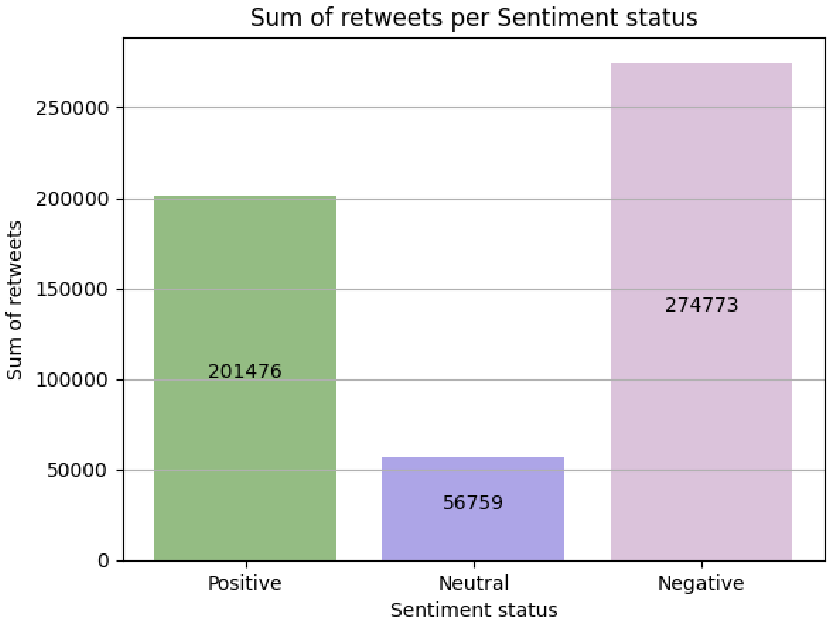


Figure 15. Sum of retweets per sentiment status in Afghanistan refugee dataset

An interesting research question for future research that can be derived from sentiment analysis is that how Afghanistan and Ukrainian refugee crises were discussed on Twitter before, during and after they occurred. The point of that would be to determine how the sentiment differs over time which would be beneficial because it indicates if the discussion is focused on assisting refugees or not. It is assumed that when the proportion of positive words increases, the topic of conversation is about helping refugees, and when the proportion of negative words increases, the topic of conversation is about the war or invasion.

Chapter 6 Bibliography

[1] "The UN Global Goals and Forced Displacement," unhcr, [Online]. Available: https://www.unhcr.org/neu/unhcr-and-sustainable-development-goals#SDG10. [Accessed 23 June 2022].

[2] "Figures at a Glance," unhcr, [Online]. Available: https://www.unhcr.org/en-ie/figures-at-a-glance.html. [Accessed 24 June 2022].

[3] S. Ellerbeck, "More people are forcibly displaced than ever before. These are the 5 things refugees need to help them find safety," weforum.org, 20 June 2020. [Online]. Available: https://www.weforum.org/agenda/2022/06/refugees-displaced-un-safety/. [Accessed 22 July 2022].

[4] M. Engler, "Germany In The Refugee Crisis – Background, Reactions And Challenges," vocaleurope, 11 May 2016. [Online]. Available: https://www.vocaleurope.eu/germany-in-the-refugee-crisis-background-reactions-and-challenges/. [Accessed 25 June 2022].

[5] I. Kaplan, "How Smartphones and Social Media have Revolutionized Refugee Migration," unhcr, 26 October 2018. [Online]. Available: https://www.unhcr.org/blogs/smartphones-revolutionized-refugee-migration/. [Accessed 26 June 2022].

[6] O. Teofilovski, "UNICEF Report Finds Half of All Refugees Are Now Children," nbcnews.com, 7 September 2016. [Online]. Available: https://www.nbcnews.com/news/world/unicef-report-finds-half-all-refugees-are-now-children-n643726. [Accessed 4 July 2022].

[7] N. WIRES, "More than one million migrants reached Europe in 2015, says UN," france24.com, 22 December 2015. [Online]. Available: https://www.france24.com/en/20151222-one-million-migrants-refugees-europe-syria-greece-2015-influx-un-war. [Accessed 5 July 2022].

[8] P. Haeck, "Global fundraiser raises over €9 billion for Ukrainian refugees," politico.eu, 9 April 2022. [Online]. Available: https://www.politico.eu/article/global-fundraiser-stand-up-for-ukraine-refugees-war-russia/. [Accessed 6 July 2022].

[9] Rebecca Hémono, Bridget Relyea, Jennifer Scott, Sinan Khaddaj, Angeliki Douka, Alison Wringe, ""The needs have clearly evolved as time has gone on.": A qualitative study to explore stakeholders' perspectives on the health needs of Syrian refugees in Greece following the 2016 European Union-Turkey agreement," *Conflict and Health,* vol. 12, no. 1, pp. 1-9, 2018.

[10] Yousef El Mourabit, Youssef El Habouz, Mustapha Lydiri, Hicham Zougagh, “A New Sentiment Analysis System Of Tweets Based On Machine Learning Approach,” *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH,* vol. 9, no. 12, pp. 41-46, 2020.

[11] b. IntelliNews, “Refugees start to flee from Ukraine to neighbouring countries,” intellinews.com, 25 Febuary 2022. [Online]. Available: https://intellinews.com/refugees-start-to-flee-from-ukraine-to-neighbouring-countries-236081/?source=bulgaria. [Accessed 8 July 2022].

[12] Kenneth R. Kaufman, Kamaldeep Bhui, Cornelius Katona, “Mental health responses in countries hosting refugees from Ukraine,” *BJPsych Open,* vol. 8, no. 3, 2022.

[13] J. Marlowe, “Refugee resettlement, social media and the social organization of difference,” *Global Networks,* vol. 20, no. 2, pp. 274-291, 2020.

[14] N. Kutscher, “The Ambivalent Potentials of Social Media Use by Unaccompanied Minor Refugees,” *Social Media and Society,* vol. 4, no. 1, p. 2056305118764438, 2018.

[15] Asma Jdaitawi, Victoria Uren, Oscar Rodriguez-Espindola, “The role of social media in promoting resilience in postdisaster recovery context: The Syrian Crisis,” *27th Annual Conference of the International Association for Management of Technology: Towards Sustainable Technologies and Innovation, IAMOT 2018,* 2018.

[16] Sandra V. Rozo, Micaela Sviatschi, “Is a refugee crisis a housing crisis? Only if housing supply is unresponsive,” *Journal of Development Economics,* vol. 148, p. 102563, 2021.

[17] Suhil M. Kiwan, Mohammad A. Alhassan, Bara M. Hamad, “Development of a simple sustainable camping shelter for addressing the needs of refugees in Jordan,” *Results in Engineering,* vol. 14, no. 100404, 2022.

[18] H. C. Smith, “'Feel the fear and do it anyway': Meeting the occupational needs of refugees and people seeking asylum,” *British Journal of Occupational Therapy,* vol. 68, no. 10, pp. 474-476, 2005.

[19] Şerif Kurtuluş, Zafer Hasan Ali Sak, Remziye Can, “Chest diseases in refugees living in a tent camp and in Turkish citizens living in the district: Ceylanpınar experience,” *Turkish Thoracic Journal,* vol. 19, no. 3, p. 117, 2018.

[20] Xavier Devictor, Quy-Toan Do, “How many years do refugees stay in exile?,” worldbank, 15 September 2016. [Online]. Available: https://blogs.worldbank.org/dev4peace/how-many-years-do-refugees-stay-exile. [Accessed 14 June 2022].

[21] K. Hodal, “Nearly half of all refugees are children, says Unicef,” theguardian.com, 7 September 2016. [Online]. Available: https://www.theguardian.com/global-

development/2016/sep/07/nearly-half-of-all-refugees-are-children-unicef-report-migrants-united-nations. [Accessed 3 July 2022].

[22] K. Fincham, “Rethinking higher education for Syrian refugees in Jordan, Lebanon and Turkey,” *Research in Comparative and International Education,* vol. 15, no. 4, pp. 329-356, 2020.

[23] J. L. McBrien, “Educational needs and barriers for refugee students in the United States: A review of the literature,” *Review of Educational Research,* vol. 75, no. 3, pp. 329-364, 2005.

[24] R. Ingram, “Refugee children are not getting the education they need,” britishcouncil.org, 2 Febuary 2016. [Online]. Available: https://www.britishcouncil.org/voices-magazine/refugee-children-not-getting-education-they-need. [Accessed 3 July 2022].

[25] Lord Clinton-Davis, Yohannes Fassil, “Health and social problems of refugees,” *Social science & medicine,* vol. 35, no. 4, pp. 507-513, 1992.

[26] Antonis A. Kousoulis, Myrsini Ioakeim-Ioannidou, Konstantinos P. Economopoulos, “Access to health for refugees in Greece: Lessons in inequalities,” *International Journal for Equity in Health,* vol. 15, no. 1, pp. 1-3, 2016.

[27] K. Kuschminder, “Deciding Which Road to Take,” *Insights into How Migrants and Refugees in Greece Plan Onward Movement,* no. 10, 2018.

[28] Asm Jdaitawi, Victoria Uren, Oscar Rodriguez-Espindola, “Communication over social media: Promoting resilience and enhancing recovery in refugee crises,” *Proceedings of the 6th European Conference on Social Media, ECSM 2019,* pp. 127-134, 2019.

[29] Rianne Dekker, Godfried Engbersen, Jeanine Klaver, Hanna Vonk, “Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making,” *Social Media and Society,* vol. 4, no. 1, p. 2056305118764439, 2018.

[30] A. Alencar, “Refugee integration and social media: a local and experiential perspective,” *Information Communication and Society,* vol. 21, no. 11, pp. 1588-1603, 2018.

[31] M. Gintova, “Understanding government social media users: an analysis of interactions on Immigration, Refugees and Citizenship Canada Twitter and Facebook,” *Government Information Quarterly,* vol. 36, no. 4, p. 101388, 2019 .

[32] Jan-Paul Brekke, Kjersti Thorbjørnsrud, “Communicating borders-Governments deterring asylum seekers through social media campaigns,” *Migration Studies,* vol. 8, no. 1, pp. 43-65, 2018.

[33] Saron Tsegaye Bekele, Stefan Stumpp, Daniel Michelis, “Influence of social media on migration and integration process,” *Proceedings of the 6th European Conference on Social Media, ECSM 2019,* pp. 29-35, 2019.

[34] P. v. Kessel, “An intro to topic models for text analysis,” medium.com, 13 August 2018. [Online]. Available: https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb. [Accessed 8 July 2022].

[35] Adina Nerghes, Ju-Sung Lee, “Narratives of the refugee crisis: A comparative study of mainstream-media and twitter,” *Media and Communication,* vol. 7, no. 2, pp. 275-288, 2019.

[36] Philip Grant, Ratan Sebastian, Marc Allassonnière-Tang, Sara Cosemans, “Topic modelling on archive documents from the 1970s: Global policies on refugees,” *Digital Scholarship in the Humanities,* vol. 36, no. 4, pp. 886-904, 2021.

[37] Muhammad Mujahid, Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Saleem Ullah, Aijaz Ahmad Reshi, Imran Ashraf, “Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19,” *Applied Sciences,* vol. 11, no. 18, p. 8438, 2021.

[38] Tom, “What is Sentiment Analysis? An Ultimate Guide for 2022,” brand24.com, 2 May 2022. [Online]. Available: https://brand24.com/blog/sentiment-analysis/#how-to-do. [Accessed 18 June 2022].

[39] Nazan Öztürk, Serkan Ayvaz, “Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis,” *Telematics and Informatics,* vol. 35, no. 1, pp. 136-147, 2018.

[40] Elizaveta Kopacheva, Victoria Yantseva, “Users’ polarisation in dynamic discussion networks: The case of refugee crisis in Sweden,” *PLoS ONE,* vol. 17, no. 2, p. e0262992, 2022 .

[41] A. Hutchinson, “New Study Shows Twitter is the Most Used Social Media Platform Among Journalists,” socialmediatoday.com, 28 June 2022. [Online]. Available: https://www.socialmediatoday.com/news/new-study-shows-twitter-is-the-most-used-social-media-platform-among-journa/626245/#:~:text=Around%20seven%2Din%2Dten%20U.S.,%25)%20and%20YouTube%20(14%25).. [Accessed 8 July 2022].

[42] N. Kossovan, “Social media makes it easy to voice an opinion without having to take personal risks or substantive action,” thestar.com, 2 March 2021. [Online]. Available: https://www.thestar.com/opinion/contributors/2021/03/02/social-media-makesit-easy-to-voice-an-opinion-without-having-to-take-personal-risks-or-substantive-action.html. [Accessed 7 July 2022].

[43] A. Tyaqi, “How to Make a Twitter Bot in Python using Tweepy,” auth0.com, 22 July 2021. [Online]. Available: https://auth0.com/blog/how-to-make-a-twitter-bot-in-python-using-tweepy/. [Accessed 10 July 2022].

[44] N. McCullum, “Building a Twitter Bot in Python with Tweepy,” nickmccullum, 11 August 2020. [Online]. Available: https://nickmccullum.com/build-twitter-bot-python-tweepy/#what-is-tweepy. [Accessed 4 July 2022].

[45] L. Hammer, “Collecting old Tweets with the Twitter Premium API and Python,” medium.com, 5 November 2019. [Online]. Available: https://medium.com/@Luca/collecting-old-tweets-with-the-twitter-premium-api-and-python-e52773d094bd. [Accessed 5 July 2022].

[46] P. Yu, “How to Access Twitter’s API using Tweepy,” towardsdatascience.com, 5 November 2019. [Online]. Available: https://towardsdatascience.com/how-to-access-twitters-api-using-tweepy-5a13a206683b. [Accessed 11 July 2022].

[47] I. A. Ogunbiyi, “Web Scraping with Python – How to Scrape Data from Twitter using Tweepy and Snscrape,” freecodecamp, 22 May 2022. [Online]. Available: Web Scraping with Python – How to Scrape Data from Twitter using Tweepy and Snscrape. [Accessed 6 July 2022].

[48] John, “A social networking service scraper in Python,” pythonawesome, 23 May 2021. [Online]. Available: https://pythonawesome.com/a-social-networking-service-scraper-in-python/. [Accessed 7 July 2022].

[49] S. Aslam, “Twitter by the Numbers: Stats, Demographics & Fun Facts,” omnicoreagency.com, 22 Febuary 2022. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/#:~:text=On%20average%2C%20500%20million%20tweets,200%20billion%20tweets%20every%20year.. [Accessed 15 July 2022].

[50] J. Sanhz, “Twitter Security Settings You Need to Change to Stay Safe,” technipages, 29 November 2021. [Online]. Available: https://www.technipages.com/twitter-security-settings-you-need-to-change-to-stay-safe. [Accessed 27 June 2022].

[51] M. Feeley, “Too much information: 4 in 5 people are still oversharing personal data on social media,” newdigitalage, 2 February 2021. [Online]. Available: https://newdigitalage.co/social-media/too-much-information-4-in-5-people-are-still-oversharing-personal-data-on-social-media/. [Accessed 27 June 2022].

[52] “What are "personal data" and when are they "processed"?,” The Data Protection Commission, [Online]. Available: https://www.dataprotection.ie/en/dpc-guidance/what-is-personal-

data#:~:text=Personal%20data%20can%20cover%20various,reasonably%20possible%20to%20find%20out.. [Accessed 8 July 2022].

[53] S. Biddiscombe, “Phishing Scams: If It Can Happen To Twitter, It Can Happen To Anyone,” forbes, 26 August 2020. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2020/08/26/phishing-scams-if-it-can-happen-to-twitter-it-can-happen-to-anyone/?sh=21b036422342. [Accessed 27 June 2022].

[54] Gatefy, “Phishing attacks use links of up to 1,000 characters,” gatefy.com, 18 March 2021. [Online]. Available: https://gatefy.com/blog/phishing-attacks-use-links-of-up-to-1000-characters/. [Accessed 9 July 2022].

[55] L. Whitney, “Phishing attack spoofs Twitter to steal account credentials,” techrepublic.com, 6 July 2020. [Online]. Available: https://www.techrepublic.com/article/phishing-attack-spoofs-twitter-to-steal-account-credentials/. [Accessed 11 July 2022].

[56] L. Irwin, “The 5 most common types of phishing attack,” itgovernance, 24 March 2022. [Online]. Available: https://www.itgovernance.eu/blog/en/the-5-most-common-types-of-phishing-attack. [Accessed 28 June 2022].

[57] J. Ellis, “Why Social Media is Increasingly Abused for Phishing Attacks,” phishlabs.com, 5 September 2019. [Online]. Available: https://www.phishlabs.com/blog/how-social-media-is-abused-for-phishing-attacks/. [Accessed 13 July 2022].

[58] T. Feinman, “Companies Need to Take Responsibility for Protecting Sensitive User Data,” entrepreneur, 2 February 2015. [Online]. Available: https://www.entrepreneur.com/article/242355. [Accessed 28 June 2022].

[59] H. N. Security, “Poor data management can cost organizations $20 million each year,” helpnetsecurity.com, 7 June 2019. [Online]. Available: https://www.helpnetsecurity.com/2019/06/07/poor-data-management/. [Accessed 15 July 2022].

[60] T. Leadership, “How to Keep Your Customers After a Data Breach,” binarynetworks.com, 13 March 2018. [Online]. Available: https://binarynetworks.com/thought-leadership/how-to-keep-your-customers-after-a-data-breach/. [Accessed 15 July 2022].

[61] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation,” *Journal of machine Learning research,* pp. 993-1022, 2003.

[62] Y. Chen, “How to generate an LDA Topic Model for Text Analysis,” yanlinc.medium.com, 17 December 2018. [Online]. Available: https://yanlinc.medium.com/how-to-build-a-lda-topic-model-using-from-text-601cdcbfd3a6. [Accessed 18 July 2022].

[63] D. Ibanez, “Topic Modeling with Latent Dirichlet Allocation,” baeldung.com, 13 July 2021. [Online]. Available: https://www.baeldung.com/cs/latent-dirichlet-allocation. [Accessed 9 July 2022].

[64] N. Seth, “Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn,” analyticsvidhya.com, 28 June 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/. [Accessed 10 July 2022].

[65] J. Rinfret, “Latent Dirichlet Allocation,” medium.com, 18 December 2019. [Online]. Available: https://medium.com/swlh/latent-dirichlet-allocation-lda-eff969bda284. [Accessed 12 July 2022].

[66] Ioana, “Latent Dirichlet Allocation: Intuition, math, implementation and visualisation with pyLDAvis,” towardsdatascience.com, 26 September 2020. [Online]. Available: https://towardsdatascience.com/latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094. [Accessed 10 July 2022].

[67] H. Naushan, “Topic Modeling with Latent Dirichlet Allocation, A practical exploration of the Natural Language Processing technique of Latent Dirichlet Allocation and its application to the task of topic modeling.,” towardsdatascience, 2 December 2020. [Online]. Available: https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8. [Accessed 29 June 2022].

[68] Yezheng Liu, Fei Du, Jianshan Sun, Yuanchun Jiang, “iLDA: An interactive latent Dirichlet allocation model to improve topic quality,” *Research Article,* vol. 4, no. 641, pp. 23-40, 2019.

[69] Jingxian Gan, Yong Qi, “Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example,” *Entropy,* vol. 23, no. 10, p. 1301, 2021.

[70] A. Omerovic, “What Is A Good Coherence Score LDA?,” wikilivre.org, 7 June 2022. [Online]. Available: https://wikilivre.org/culture/what-is-a-good-coherence-score-lda/. [Accessed 16 July 2022].

[71] W. Guntermann, “What is perplexity in topic modeling?,” answersblurb.com, 13 April 2022. [Online]. Available: https://answersblurb.com/what-is-perplexity-in-topic-modeling. [Accessed 18 July 2022].

[72] A. CR, “Topic Modeling using Gensim-LDA in Python,” medium.com, 26 July 2020. [Online]. Available: https://medium.com/analytics-vidhya/topic-modeling-using-gensim-lda-in-python-48eaa2344920. [Accessed 18 July 2022].

[73] R. Reed, “Everything You Need To Know About Emoji,” smashingmagazine.com, 14 November 2016. [Online]. Available: https://www.smashingmagazine.com/2016/11/character-sets-encoding-emoji/. [Accessed 20 July 2022].

[74] D. Deshpand, “Validating gensim’s topic coherence pipeline,” rare-technologies.com, 18 July 2016. [Online]. Available: https://rare-technologies.com/validating-gensims-topic-coherence-pipeline/. [Accessed 27 July 2022].

[75] B. Mabey, “Python library for interactive topic model visualization. Port of the R LDAvis package,” zzun.app, 16 January 2022. [Online]. Available: https://zzun.app/repo/bmabey-pyLDAvis. [Accessed 20 July 2022].

[76] U. Nation, “Goal 5: Achieve gender equality and empower all women and girls,” un.org, [Online]. Available: https://www.un.org/sustainabledevelopment/gender-equality/. [Accessed 10 August 2022].

[77] T. U. R. Agency, “UNHCR policy on,” unhcr.org, September 2009. [Online]. Available: https://www.unhcr.org/en-ie/protection/hcdialogue%20/4ab356ab6/unhcr-policy-refugee-protection-solutions-urban-areas.html. [Accessed 14 August 2022].

[78] R. Chadwick, “Tweepy for beginners,” towardsdatascience.com, 1 July 2019. [Online]. Available: https://towardsdatascience.com/tweepy-for-beginners-24baf21f2c25. [Accessed 25 July 2022].

[79] J. Graber, “Python Friday #116: Search Twitter from Tweepy,” improveandrepeat.com, 1 April 2022. [Online]. Available: https://improveandrepeat.com/2022/04/python-friday-116-search-twitter-from-tweepy/. [Accessed 25 July 2022].

[80] Z. Motti, “Geolocating tweets via spatial inspection of information inferred from tweet meta-fields,” *International Journal of Applied Earth Observation and Geoinformation,* vol. 105, p. 102593, 2021.

[81] Marcio Pereira Basilio, Valdecy Pereira, Antonio Fernandes Costa Neto, ax William Coelho Moreira de Oliveira, Orlinda Claudia Rosa de Moraes, Samya Cotta Brand~ao Siqueira, “Plate notation for LDA with a Dirichlet distribution,” researchgate, March 2021. [Online]. Available: https://www.researchgate.net/figure/Plate-notation-for-LDA-with-a-Dirichlet-distribution_fig1_349899227. [Accessed 5 July 2022].

[8 P. Ninja, “New Twitter Phishing Campaign Targets Verified Accounts,” Privacy Ninja, 6
2] December 2021. [Online]. Available: https://www.privacy.com.sg/cybersecurity/new-twitter-phishing-campaign-targets-verified-accounts/. [Accessed 2 July 2022].

Appendix A – List of Abbreviation

Application Programming Interface
(API) iv
Comma-Separated Values
(CSV) 23
Global Positioning System
(GPS) 55
Graphics Processing Unit
(GPU) 55
Latent Dirichlet Allocation
(LDA) iv
Sustainable Development Goals
(SDG) iv
United Nation
(UN) iv
United Nations High Commissioner for Refugees
(UNHCR) 11
User Interface
(UI) 23

Appendix B – Code Repository

https://github.com/Misbah-Rizaee/Research-Project

Appendix C – Screenshots

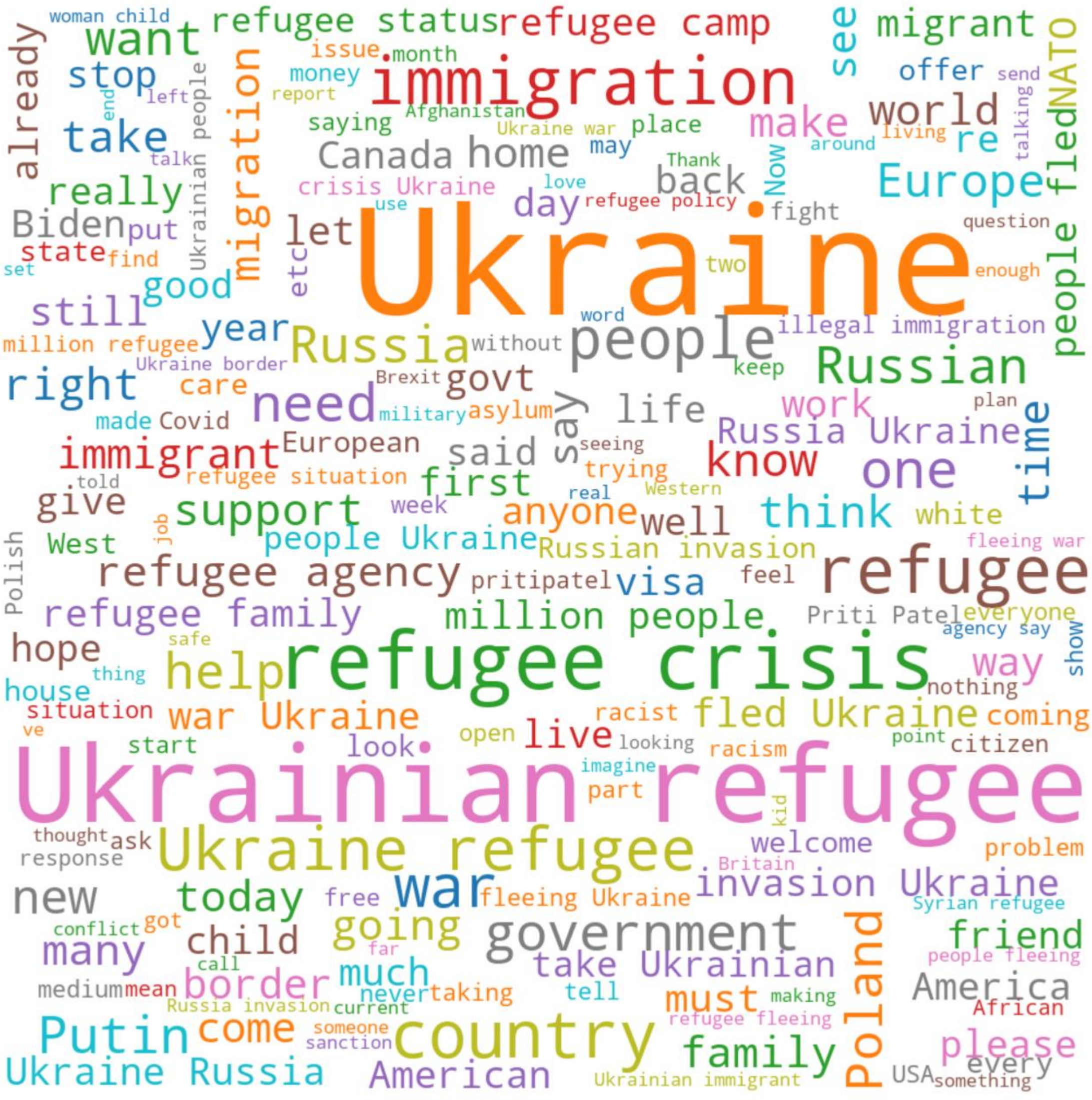


Figure 16. WordCloud visualization of Ukrainian refugee dataset

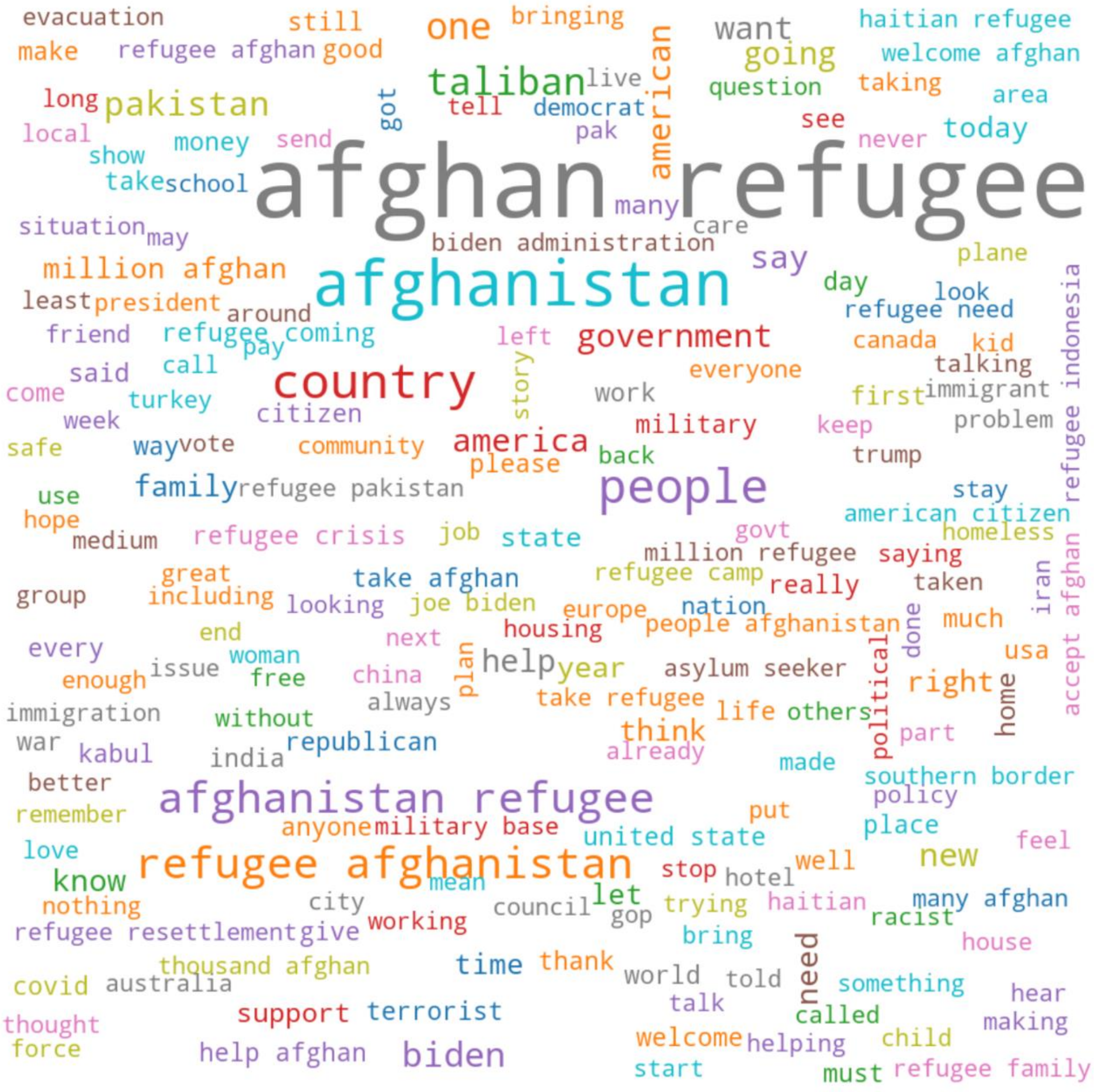
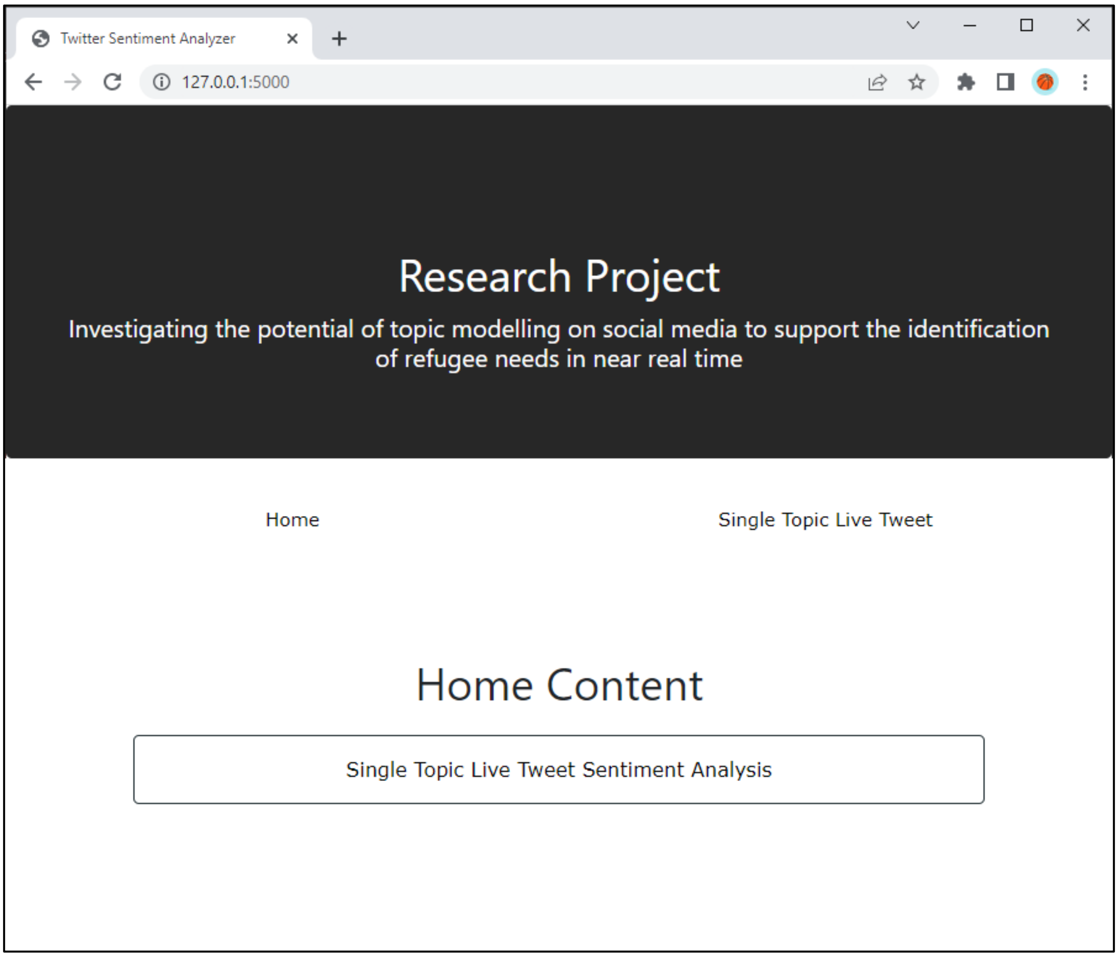


Figure 17. WordCloud visualization of Afghanistan refugee dataset


Twitter Sentiment Analyzer
127.0.0.1:5000

Research Project

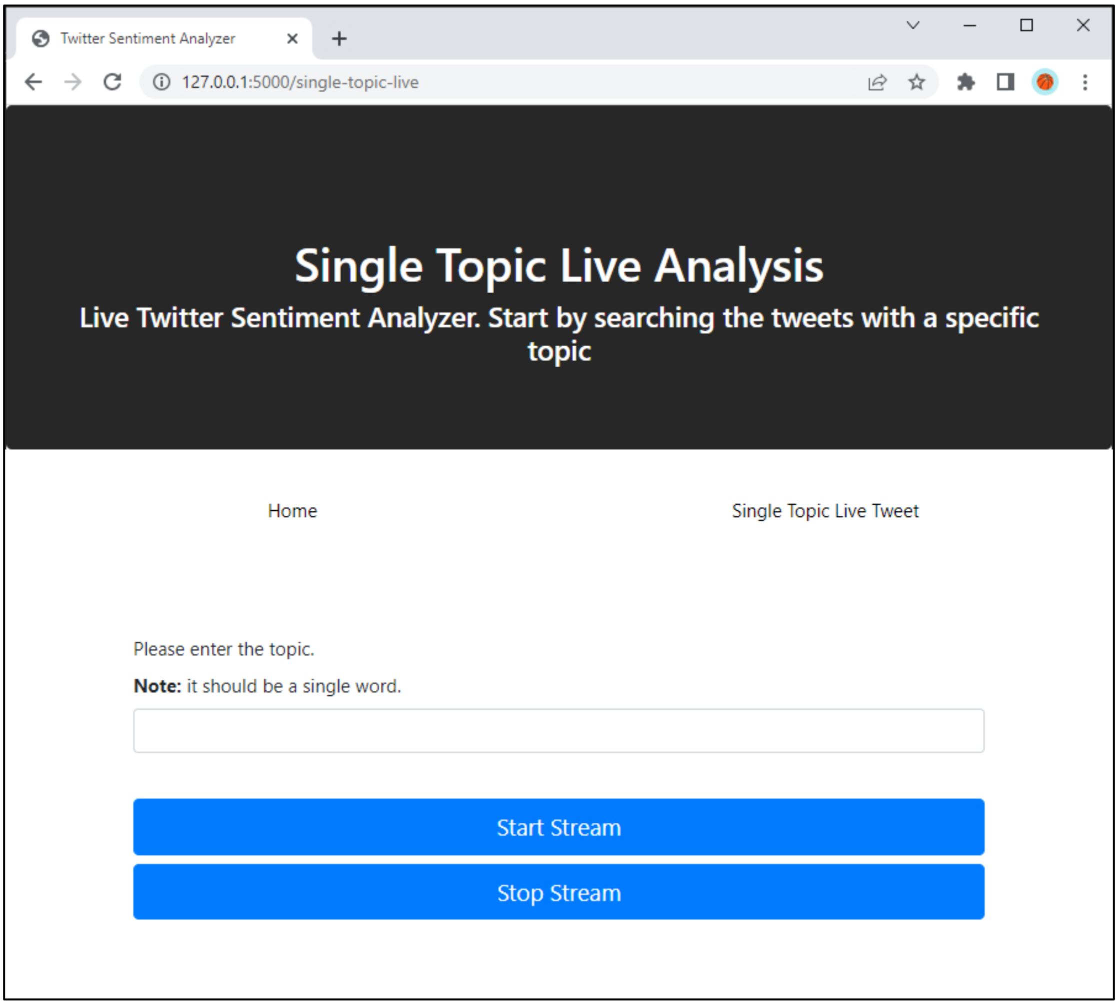
Investigating the potential of topic modelling on social media to support the identification of refugee needs in near real time

Home
Single Topic Live Tweet

Home Content

Single Topic Live Tweet Sentiment Analysis

Figure 18. Screenshot of the main page in the user interface



Twitter Sentiment Analyzer

127.0.0.1:5000/single-topic-live

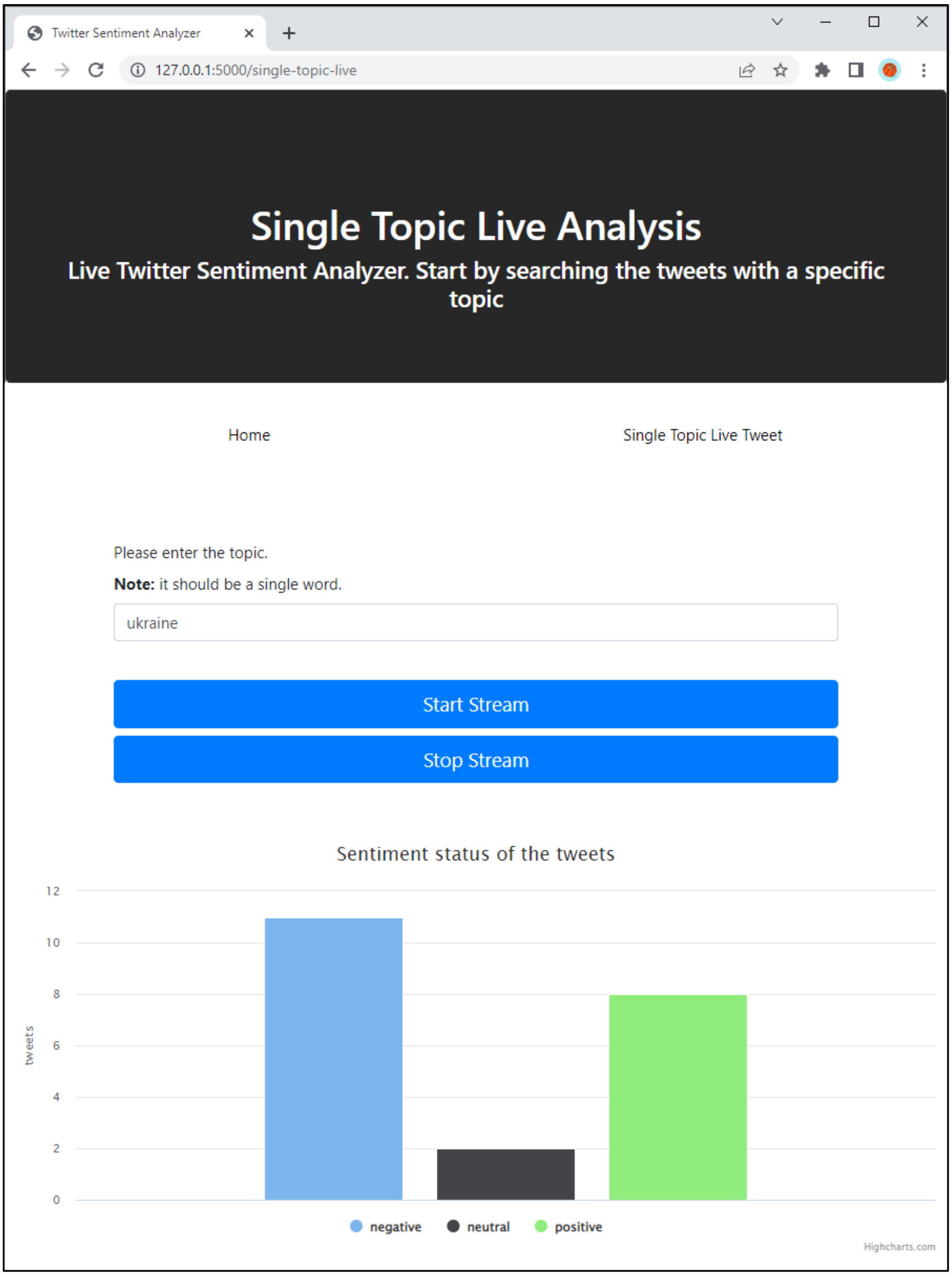
Single Topic Live Analysis

Live Twitter Sentiment Analyzer. Start by searching the tweets with a specific topic

Please enter the topic.

Note: it should be a single word.

Figure 19. Screenshot of the second page in the user interface. This is a live Twitter Sentiment analyser



Single Topic Live Analysis

Live Twitter Sentiment Analyzer. Start by searching the tweets with a specific topic

Please enter the topic.

Note: it should be a single word.

ukraine

Sentiment status of the tweets

Figure 20. The live Twitter Sentiment analyser is started and the chart is showing the sentiment of the live tweets which are being collected