FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

# ANALYZING FOOD TRENDS ON TWITTER

Word count: 11499

Student number : 01810449

Supervisor: Prof. Dr. Dirk Van den Poel

Master's Dissertation submitted to obtain the degree of:

Master in Business Engineering: Data Analytics

Academic year: 2022-2023

UNIVERSITEIT GENT

# Confidentiality of the master's dissertation

The author and the promotor give permission to use this master's dissertation for consultation and to copy parts of it for personal use. Every other use is subject to the copyright laws, more specifically the source must be extensively specified when using results from this master's dissertation.

I declare that the research was conducted in accordance with the rules governing scientific and academic integrity. I have read, and acted in accordance with, the Code of Ethics of the Faculty.

Arthur Galoppin

# Table of contents

1. Abstract ..... 1
2. Introduction ..... 1
2.1. Research questions ..... 1
2.2. Relevance ..... 2
3. Literature review ..... 3
3.1. Social representations ..... 3
3.2. Geolocated communities ..... 3
3.3. Geolocated communities and their attitudes ..... 4
3.4. Social representations shaping the attitudes ..... 4
4. Methodology ..... 5
4.1. Data collection ..... 6
4.1.1. Retrieve Tweets with Twitter API ..... 6
4.1.2. Filter Tweets ..... 8
4.2. Geolocated communities ..... 9
4.2.1. Location pre-processing ..... 9
4.2.2. Clustering ..... 11
4.2.2.1. DBSCAN ..... 11
4.2.2.2. Algorithm ..... 11
4.2.2.3. Parameter selection ..... 14
4.3. Geolocated communities and their attitudes ..... 15
4.3.1. Sentiment analysis ..... 16
4.3.1.1. Twitter-XLM-roBERTa-base sentiment model ..... 17
4.3.1.2. Language model for text embeddings ..... 18
4.3.1.3. Fine tuning for sentiment analysis ..... 19
4.3.2. Sentiment per community ..... 21
4.4. Social representations shaping the attitudes ..... 21
4.4.1. Topic modelling ..... 22
4.4.1.1. Generating text embeddings ..... 24
4.4.1.2. Clustering embeddings ..... 25
4.4.1.3. Most important words ..... 26
4.4.2. Sentiment per topics ..... 26
5. Results ..... 26
5.1. Geolocated communities ..... 26
5.2. Geolocated communities and their attitudes ..... 27
5.3. Social representations shaping attitudes ..... 28
6. Limitations ..... 30
7. Conclusion ..... 31

# List of figures

Figure 1 Research methodology diagram ..... 5
Figure 2 Research methodology diagram: Data Collection ..... 6
Figure 3 Research methodology diagram: Geolocated Communities ..... 9
Figure 4 Distance matrix ..... 11
Figure 5 Core and non-core Twitter users ..... 12

Figure 6 Directly density-reachable Twitter users ..... 12
Figure 7 Density-reachable Twitter users ..... 12
Figure 8 Density connected Twitter users ..... 13
Figure 9 Core, border and noise Twitter users ..... 13
Figure 10 k-distance plot (k=20) ..... 14
Figure 11 Research methodology diagram: Geolocated Communities And Their Attitudes ..... 15
Figure 12 Sentiment analysis: machine learning approach ..... 17
Figure 13 Feature extraction from simplified neural network architecture of language model ..... 19
Figure 14 Fine-tuning language models for sentiment analysis ..... 20
Figure 15 Research methodology diagram: Social Representations Shaping The Attitudes ..... 22
Figure 16 Word embeddings in 3-dimensional vector space ..... 24
Figure 17 Fine-tuning of language model for better word embeddings ..... 25
Figure 18 Geolocated communities on Twitter ..... 27
Figure 19 World cultural regions ..... 27
Figure 20 Visualization of topic clusters in 2 dimensions ..... 30

# List of tables

Table 1 Hashtags and related food trends ..... 7
Table 2 Dataframe structure of retrieved Tweets ..... 8
Table 3 Tweet text and relation to food trend ..... 9
Table 4 Twitter users and their coordinates ..... 10
Table 5 Preview of sentiment analysis on Tweets ..... 21
Table 6 Sentiment per community ..... 28
Table 7 Sentiment per topic ..... 29

# 1. Abstract

This master's dissertation aims to extract valuable insights from raw textual Twitter data about food trends for various stakeholders such as food-related businesses, policymakers, and consumers who are affected by these trends. The study's main contribution is its methodology of analysis, which can be applied to any food trend or trend in general at any time, since food trends change over time. The research uses the DBSCAN clustering algorithm to identify geolocated communities of Twitter users around the world who talk about the same hashtag. The study also employs pretrained language models to conduct sentiment analysis and topic modelling on tweets in multiple languages. Insights from clustering, sentiment analysis and topic modelling are integrated to explore differences in the sentiment of tweets across geolocated communities as well as differences in sentiment depending on the topic covered in the tweets. The results are interesting, but, what is especially surprising is how well the state of the art natural language processing (NLP) techniques perform for sentiment analysis and topic modelling, handling sarcasm, context-dependent sentiments, humour and slang with ease.

# 2. Introduction

Food trends have been around for centuries, with early examples including the use of spices to flavour food. The ancient Egyptians, for instance, were known for their use of spices like mint, coriander, cinnamon, and onion, which were often used for their flavour and medicinal properties ("History of Spices," n.d.). Today, food trends continue to evolve, but the speed at which they spread and the level of global influence they have has increased in recent years. The rise of social media and the ability to share information instantly has greatly contributed to the rapid spread of food trends around the world. All the data generated through these interactions on social media is stored online, creating a vast digital repository that can be analyzed to gain socially or scientifically relevant insights.

## 2.1. Research questions

This research study aims to retrieve valuable insights about food trends by analysing the global conversation surrounding them on Twitter. In order to do this in a structured way, the study attempts to answer several research questions. It is important to notice the logical sequence of the research questions. Each subsequent research question integrates the findings of the previous question.

The first research question seeks to unveil distinct **geolocated communities** of Twitter users around the world. These geolocated communities can be seen as groupings of a significant amount

of users that live in a relatively close spatial proximity to each other. The application of clustering techniques allows for the discovery of such geolocated communities. Geographical location is a relevant factor to consider when studying food trends, as food traditions differ widely depending on location.

The second research question examines how Twitter users engage with the food trends on Twitter. Specifically, the study aims to determine whether people have positive or negative **attitudes** towards certain trends. By performing sentiment analysis on the Tweets, the research can uncover these attitudes that people hold towards different food trends. This in combination with the findings related to the first research question allows us to see how the attitudes towards food trends vary across different geolocated communities.

The third research question explores the underlying drivers for people to be involved with a certain food trend. Sentiment analysis can reveal the attitudes that people hold towards a certain food trend. However, by delving deeper, the study aims to reveal the underlying knowledge, beliefs, and ideas that influence and shape these attitudes . These underlying drivers, also known as **social representations,** can be uncovered by doing topic modelling on the content of the Tweets. If this insight is combined with the findings related to the second research question, the social representations shaping people's positive or negative attitudes towards food trends can be revealed.

After elaborating on all research questions insights are obtained in how social media can be leveraged to promote food practices in a certain geolocated community. For example, are people who live in a certain geolocated community sharing their positive experiences with a particular food trend because they enjoy the taste, appreciate the health benefits, or simply because it's a popular trend that they want to be associated with? Similarly, what motivates people to criticize or express negative sentiments towards a particular food trend? Is it due to concerns about the environmental impact, ethical considerations, or health risks associated with the trend?

## 2.2. Relevance

The ability to provide valuable insights into food trends can be of interest for various stakeholders. By analysing the trends and discussions that emerge on the social media platform, the study can help inform food-related businesses, policymakers, and consumers who are directly impacted by these trends.

However, the significance of this research study extends beyond its ability to provide valuable insights into food trends for various stakeholders. In addition, the study also holds value in terms of its methodology of analysis as it is an extension to the current existing literature. The analytical

approach employed in this research can also be generalized and applied to any food trend or trend in general other than the ones that were used for the study.

# 3. Literature review

In this literature review, the prior research, related to the established research questions, is summarised and analyzed critically. Also, gaps in knowledge are identified, and suggestions of potential avenues for future research are made.

## 3.1. Social representations

In social psychology, social representation refers to the shared beliefs, ideas, and knowledge held by a particular community about certain subjects. These shared beliefs, ideas, and knowledge are responsible for shaping people's attitudes towards these subjects. Social representations are established trough social interactions within a community and guide individuals in their decision-making (Moscovici, 1984). For instance, in Northern China people share the believe that eating dog meat promotes warmth in the body when consumed. This social representation of eating dog meat influences their positive attitudes towards including dog meat in their diet during winter months (Qin, 2016). On the other hand, in both North and South Korea the consumption of dog meat is believed to keep up your stamina, specifically during the summer. This social representation shapes positive attitudes towards eating dog food during summer months (Whan-woo, 2019).

Various studies with surveys already indicated that social representations have a significant impact on people's willingness to try new foods (Bäckström, Pirttilä-Backman, & Tuorila, 2004; Bartels & Reinders, 2010; Huotilainen, Pirttilä-Backman, & Tuorila, 2006; Onwezen & Bartels, 2013). The collaboratively created and shared knowledge about unfamiliar foods helps people to cope with the novelty, which is one reason why social representations are so important for determining peoples' willingness to try new foods (Huotilainen et al., 2006; Moscovici, 1984).

Building upon the definition of a social representation, one can understand that geolocated communities offer more than just geographical data. They could also provide insights into shared beliefs, ideas, and knowledge, shaping the attitudes of the people within those communities (Moscovici, 1984; Stefanidis, Crooks, & Radzikowski, 2013). In subsection 3.2 there is elaborated on prior research that tries to reveal geolocated communities on Twitter. In subsection 3.3 there is elaborated on prior research that tries to reveal the attitudes of geolocated communities on Twitter. In subsection 3.4 there is elaborated on prior research that tries to reveal the social representations that can shape these attitudes on Twitter.

## 3.2. Geolocated communities

To discover geolocated communities of Twitter users and their distribution around the world, the ability to retrieve the location of the Twitter user who posted the Tweet is a necessary requirement. Using Twitters API it is possible to extract Tweets that includes both the Tweet text and relevant metadata. This metadata can provide various details about the Tweet, such as the author's username, the post's timestamp, and potentially the author's location. Harvesting this information can be done with relative ease. When it became possible to easily trace the location from which a Tweet originates, an increasing number of studies have emerged that attempt to leverage this information, for example, using clustering algorithms to detect geolocated communities (Bakillah, Li, & Liang, 2015; Gao et al., 2017; Stefanidis et al., 2013).

## 3.3. Geolocated communities and their attitudes

After detecting geolocated communities, it is possible to determine the attitudes of these communities by applying sentiment analysis on the content they share (Deitrick & Hu, 2013; Pang & Lee, 2008). As a result, an increasing number of Twitter-based studies are adopting a combined approach of identifying geolocated communities using clustering techniques and analyzing the sentiment of the content shared within them to understand public opinion dynamics. For example, Hridoy, Ekram, Islam, Ahmed, & Rahman (2015) used the DBSCAN clustering algorithm and sentiment analysis to analyze public opinion on the iPhone 6 across the USA. Similarly, Stojanovski, Strezoski, Madjarov, Dimitrovski, & Chorbev (2018) employed DBSCAN clustering and sentiment analysis on Twitter messages related to the 2014 FIFA World Cup to examine the emotional attitudes of fans during the event and identify social hotspots in New York.

Despite an increasing amount of Twitter based studies adopting this combined approach, previous research on social media in a food related context has mainly focused on content analysis, with limited exploration of the spatial dimension (Pindado & Barrena, 2020). Nevertheless, there are notable exceptions in the field. For instance, Pindado and Barrena (2020) used a database of tweets about new foods to identify geolocated communities using DBSCAN clustering and compared the attitudes expressed across these communities with sentiment analysis.

## 3.4. Social representations shaping the attitudes

Recall that social representations as the shared beliefs, ideas, and knowledge that shape the attitudes of people within a community about a certain subject. The previous section was devoted to prior studies that combine the detection of geolocated communities and sentiment analysis to explore different attitudes across these communities. This section focuses on prior research that tries to reveal the social representations which could be the underlying drivers for certain attitudes.

There are a lot of studies out there that do topic modelling to uncover the main topics in a corpus of Tweets. Topic modelling can be used to reveal the shared ideas, believes and knowledge about a certain subject. However, this literature review didn't succeed in finding a source that combines topic modelling with sentiment analysis to see how the different underlying topics relate to positive or negative attitudes. Most studies looked at sentiment analysis and topic modelling as separate concepts. As such, the assumption is made that there is no to little research done that integrates sentiment analysis and topic modelling providing an opportunity for further investigation.

# 4. Methodology

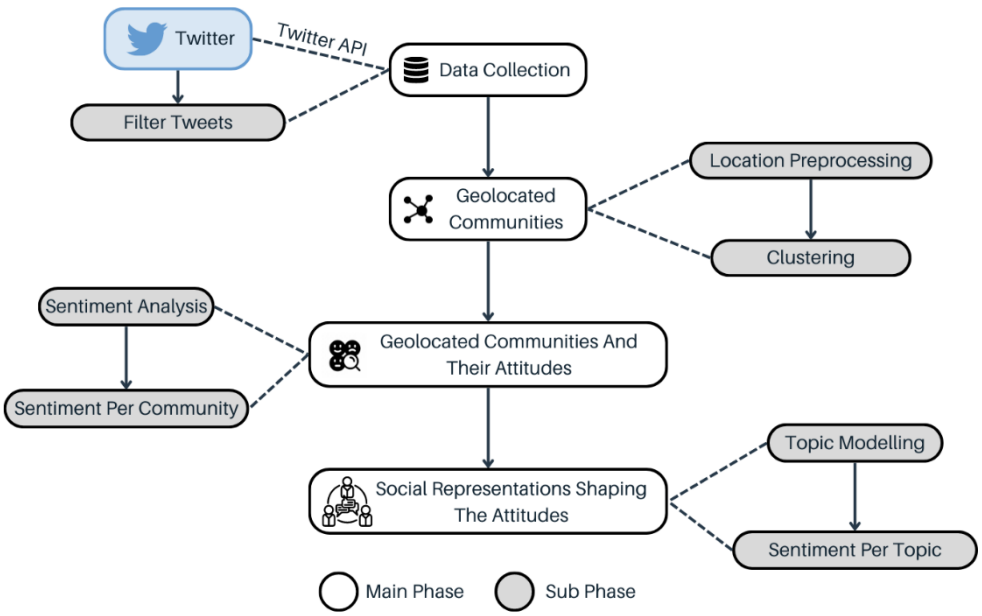


*Figure 1 Research methodology diagram*

Figure 1 is a research methodology diagram that serves as a schematical overview of the phases that will be executed in this research, and the relationships between them. There are two types of phases and two types of relationships that are distinguished. The main phases are represented as white fields with a black border and the subphases are represented by grey fields with a black border. A sequence relationship between two phases indicates a logical order between two phases, and is represented by a solid black arrow. A decomposition relationship indicates a hierarchical structure between two phases, and is represented by a dotted line.

For the methodology of this research, four main phases are distinguished. Except for the first main phase, data collection, which is a prerequisite to do any further analysis, each main phase is related to a research question. As pointed out earlier, the research questions follow up in a logical sequence where each subsequent research question integrates the findings of the previous research question. As result, all the main phases follow up in a sequential order. Each of the main phases on their turn

can be decomposed into substages which also have a sequential order between them. This section provides a detailed description of each phase that was carried out. It only focuses on the activities performed at each phase, while the results obtained are later presented in section 5.

Together, Jupyter Notebook and Python are used to execute the different phases of research. Jupyter Notebook is a web-based interactive computational environment that supports several programming languages including Python. Python is a high-level programming language that is widely used in data science and machine learning. All the datafiles containing the Tweets that were retrieved as well as the code to analyze this data can be found on GitHub in the public repository that can be found by clicking on following link.

## 4.1. Data collection

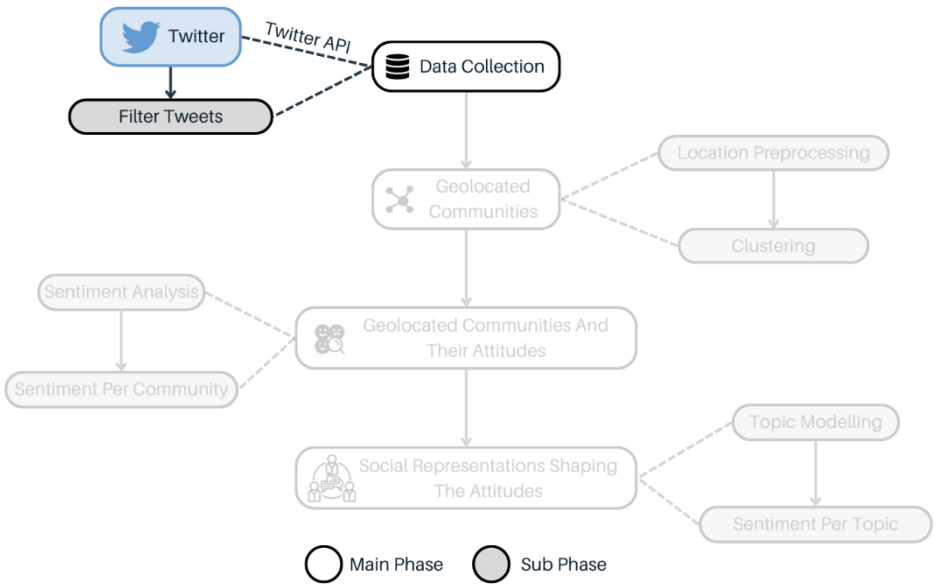


*Figure 2 Research methodology diagram: Data Collection*

This subsection explains how to collect and filter tweets for analysis. It describes the methods for gathering tweets and removing the ones that are not relevant.

### 4.1.1. Retrieve Tweets with Twitter API

Twitter hashtags are keywords or phrases preceded by the # symbol, without spaces or punctuation. Twitter users can include hashtags in their post to categorize their Tweet according to the subject that is discussed in the Tweet. For example, #COVID19 is a popular hashtag used in Tweets discussing the pandemic caused by the coronavirus or #MeToo is often mentioned in Tweets that aim to raise awareness about sexual harassment and assault. Hashtags make it easy for other Twitter users to find content that interests them because they can retrieve groups of Tweets around their subjects of interest by searching for the hashtags associated with these subjects. This means

that including a hashtag related to a certain subject can increase the visibility of the Tweet among the audience interested in that subject.

As a starting point for this research, it is necessary to identify key food trends discussed on Twitter and their corresponding hashtags. Table 1 provides an overview of the food trends identified for this research and their corresponding hashtags on Twitter.

| Hashtag | Food trend |
|---|---|
| #vegan | Veganism, which is a lifestyle and diet that excludes all animal products and animal by-product. |
| #vegetarian | Vegetarianism, which is a diet that excludes meat and fish but may include other animal products such as eggs and dairy. |
| #paleo | Paleolithic diet, which is a diet based on the types of foods presumed to have been eaten by early humans. |
| #plantbased | Plant-based diet, which is a diet that focuses on foods derived from plants, including vegetables, fruits, whole grains, legumes, nuts and seeds. |
| #healthylifestyle | The trend of pursuing a healthy lifestyle through balanced nutrition, regular exercise, and other healthy habits. |

*Table 1 Hashtags and related food trends*

The value of this research lies not in the identified food trends themselves, but rather in the methodology of analysis that can be generalized to any food trend beyond those used in this study. As such, the choice of food trends was somewhat arbitrary or based on personal interests. For the pursuit of the research, the applied methodology and obtained results are only shown for the food trend about veganism. The GitHub repository contains numerous tweets for each hashtag listed in Table 1. Therefore, to explore results for other food trends, you can execute the Analysis.ipynb file from the GitHub repository using a different hashtag.

To retrieve Tweets related to the identified food trends, the Twitter API can be used. This tool, provided by Twitter, allows developers to access and interact with Twitter's platform and data. With the Twitter API, you can retrieve Tweets, user profiles, trends, and other information. For this research, the API was used to retrieve the text of Tweets containing the hashtags as well as metadata linked to the Tweets such as publication date and time, language, author's username and location, etc.

All retrieved Tweets related to a certain food trend are stored in a dataframe where each Tweet occupies a separate row and columns are included for the publication date and time, the author's username, the text of the Tweet, the author's location, and the language of the Tweet. Table 2 shows a part of the dataframe related to veganism, serving as an example to capture the structure.

| datetime | author | text | location | lang |
|---|---|---|---|---|
| 2023-03-12 14:29:05 | Vink6741 | @MEGroenstege Gisteren bij #kassa. Diverse #vegan burgers getest. Eindconclusie: rubber en karton smaak. Het was de saus en broodje die het net eetbaar maakte. 1 Mac burger kreeg een voldoende, de rest een dikke onvoldoende. Noem het geen vlees, maar groente of #soja burgers. Nep vlees is het. | Netherlands | nl |
| 2023-03-13 13:59:29 | layHunterr | who is behind the sudden fall of 2 banks in US? #competition #influencer #influencermarketing #fridayfeeling #MondayMotivation #tbt #traveltuesday #vegan #fitness #FIFAWorldCup #Qatar2022 #TikTok #Argentina #Messi #WorldCup #subscriber #modelmodeling #viral | India | en |
| 2023-03-13 13:57:07 | YT_AltBattles1 | Vegans who are animal rights activists are complete NPCs. #animalrightsactivists #vegan #animalrights #npcs #npc2023 | Calgary, Canada | en |
| 2023-05-23 06:31:06 | 9tomoon | la viande, le poisson, les oeufs, les produits laitiers... Tout ce qui provient d'un animal mdr et ça depuis 12 ans #vegan https://t.co/YnLk5lWh2v | animal protection&vegan 🐾 | fr |
| 2023-03-13 13:55:36 | avon_bradford | 🥑 Millions swear by going vegan as a remedy for skin problems, low energy and so much more. 🥗 Trying to jump on the plant based train? 🚆 Here's some helpful advice from those who have already succeeded! https://t.co/mYLQ90LkBn #PlantBased #Vegan #HealthyLiving https://t.co/RF85XjyuyR | Bradford, England | en |
| 2023-03-13 09:05:01 | djventilator | Ik ben niet #vegan, omnivoor of carnivoor, geen vegetariër of flexitariër. Labels suggereren dat ik geen keuze heb, dat ik niet anders kan. Ik kies er voor om #duurzaam te leven, zo eet ik meestal plantaardige eiwitten. Dit doe ik bewust en in vrijheid. https://t.co/FjQBdd5moW | Planet Earth | nl |

*Table 2 Dataframe structure of retrieved Tweets*

### 4.1.2. Filter Tweets

After retrieving Tweets and storing them in a dataframe, it is essential to filter out Tweets that are not relevant for analysis. The filtering process involves several steps. First, retweets are removed by checking if the text of the Tweet begins with 'RT.' Next, duplicate Tweets are removed by checking for identical text and retaining only the first occurrence.

The purpose of retrieving Tweets that include the hashtag #vegan is to obtain Tweets that discuss the subject of veganism. However, it has been observed that many Tweets discuss entirely different subjects while still including the hashtag #vegan. This common misuse of Twitter hashtags occurs because users aim to drive traffic or attention to their Tweets from a wider audience by using excessive, unrelated hashtags in a single Tweet.

To filter out unrelated Tweets, a pre-trained multilingual language model fine-tuned for zero-shot text classification can be utilized. This model takes a text and a list of topics as input and assigns a probability to each topic indicating the likeliness that the text belongs to that topic. Only one input topic, "food", was selected, and the input text was the Tweet with the #vegan hashtag removed to avoid bias. Notice that "veganism" could also have been chosen as input topic or "food" together with "veganism", but, we are already satisfied if the Tweet covers the subject food in general. So for this case, the model outputs scores between 0 and 1 based on the degree to which a Tweet talks about food. The threshold against which the probability score is evaluated to determine whether to keep or discard the Tweet is set at 0.5. This approach has been frequently employed to classify text according to the topic it covers (Mishra, 2022). However, using it to filter Tweets to ensure they are relevant to the hashtag is an approach proposed by this study that is not based on any prior work. It has been found to be highly accurate (table 3). The details of how such a language model is

constructed are not discussed here but are explained in more detail in section 4.3 and 4.4 on sentiment analysis and topic modelling respectively.

| text | abouthashtag | keep |
|---|---|---|
| @MEGroenstege Gisteren bij #kassa. Diverse #vegan burgers getest. Eindconclusie: rubber en karton smaak. Het was de saus en broodje die het net eetbaar maakte. 1 Mac burger kreeg een voldoende, de rest een dikke onvoldoende. Noem het geen vlees, maar groente of #soja burgers. Nep vlees is het. | 0.997171 | yes |
| who is behind the sudden fall of 2 banks in US? #competition #influencer #influencermarketing #fridayfeeling #MondayMotivation #tbt #traveltuesday #vegan #fitness #FIFAWorldCup #Qatar2022 #TikTok #Argentina #Messi #WorldCup #subscriber #modelmodeling #viral | 0.000315 | no |
| Vegans who are animal rights activists are complete NPCs. #animalrightsactivists #vegan #animalrights #npcs #npc2023 | 0.952235 | yes |
| la viande, le poisson, les oeufs, les produits laitiers... Tout ce qui provient d'un animal mdr et ça depuis 12 ans #vegan https://t.co/YnLk5lWh2v | 0.998932 | yes |
| 🥦 Millions swear by going vegan as a remedy for skin problems, low energy and so much more. 🥗 Trying to jump on the plant based train? 🧵 Here's some helpful advice from those who have already succeeded! https://t.co/mYLQ90LkBn #PlantBased #Vegan #HealthyLiving https://t.co/RF85XjyuyR | 0.697887 | yes |
| Ik ben niet #vegan, omnivoor of carnivoor, geen vegetariër of flexitariër. Labels suggereren dat ik geen keuze heb, dat ik niet anders kan. Ik kies er voor om #duurzaam te leven, zo eet ik meestal plantaardige eiwitten. Dit doe ik bewust en in vrijheid. https://t.co/FjQBdd5moW | 0.979008 | yes |

*Table 3 Tweet text and relation to food trend*

## 4.2. Geolocated communities

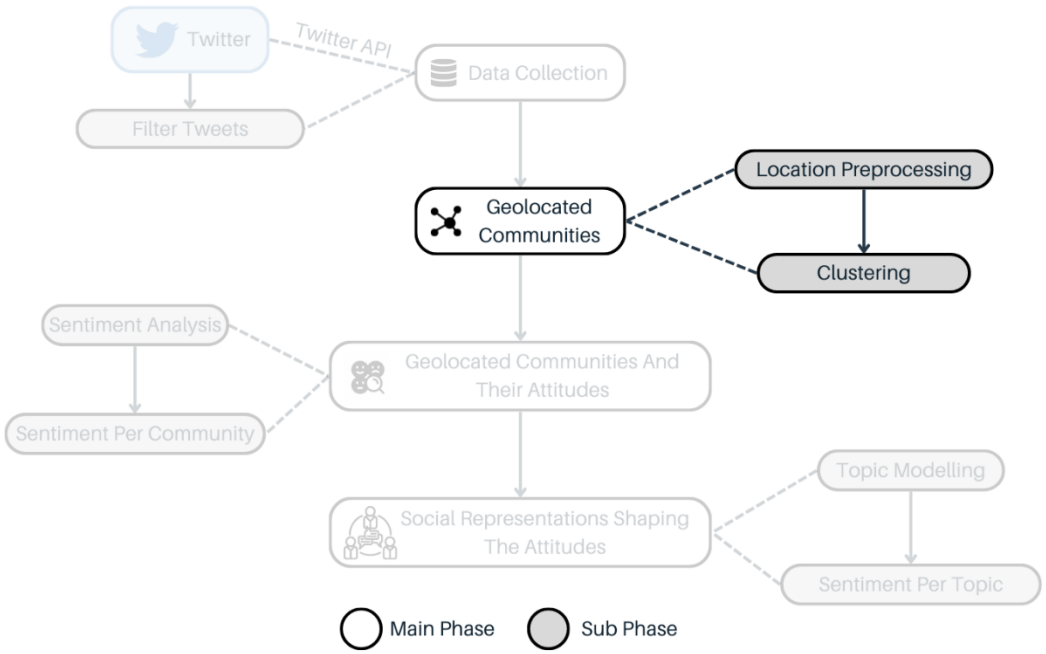


*Figure 3 Research methodology diagram: Geolocated Communities*

This subsection explains how to detect geolocated communities of Twitter users who shared tweets about #vegan. First there is looked at how the distances between users are derived. Then there is zoomed into the algorithm used to cluster them together based on their distances from one another.

### 4.2.1. Location pre-processing

To group Twitter users into geolocated communities, the pairwise distances between every unique pair of Twitter users must be calculated and provided as input to the clustering algorithm. The

process to derive the distances between every unique pair of Twitter users involves two steps. First, a new dataframe of unique Twitter users with their coordinates must be derived from the current dataframe of Tweets. Second, from this dataframe of unique Twitter users with their coordinates, a matrix containing the distances between every unique pair of Twitter users must be derived. The details of these two steps are explained in the following two paragraphs respectively.

Starting from a dataframe of Tweets, the first step in the analysis is to derive a new dataframe in which Tweets are aggregated by user and retain only the user and location columns. This ensures that each user appears only once in the dataframe, regardless of the number of tweets they made about #vegan. Next, all Twitter users who did not provide location information on Twitter are removed from the dataset. It was observed that some users entered only their country as their location, without providing more specific information about their city or street. This resulted in the assignment of coordinates somewhere in the middle of the country, which could be a sparsely populated area. Empirical evidence showed that this was problematic, as separate clusters were created in the middle of Canada and Australia, areas with very low population density. To address this issue, for users who entered only their country as their location, their location was changed to the capital city of that country. The location column now contains locations that were either manually entered by the users or slightly adapted if they only entered their country. A function is then applied to extract coordinates from these manually entered locations. Twitter users for whom coordinates could not be found are subsequently removed from the dataset. The resulting dataframe contains one row per unique user, with columns for their username and coordinates (Table 4).

| author | coordinates |
|---|---|
| Vink6741 | [52.2434979, 5.6343227] |
| YT_AltBattles1 | [51.0460954, -114.065465] |
| avon_bradford | [53.7944229, -1.7519186] |
| djventilator | [14.3601126, 100.5772104] |

*Table 4 Twitter users and their coordinates*

Based on the obtained dataframe, the distance between every pair of Twitter users should be derived. Distance here is defined as shortest distance between Twitter users on the surface of the globe also known as great-circle distance. The haversine formula is used to determine the great-circle distance between two points on the globe given their longitudes and latitudes (“Haversine formula,” n.d.). This results in a distance matrix that contains the pairwise great-circle distances, in kilometres, between the users (figure 4).

| | Vink6741 | YT_AltBattles1 | avon_bradford | djventilator |
|---|---|---|---|---|
| Vink6741 | 0.0 | 7217.3 | 523.0 | 9080.6 |
| YT_AltBattles1 | 7217.3 | 0.0 | 6770.6 | 12003.6 |
| avon_bradford | 523.0 | 6770.6 | 0.0 | 9510.5 |
| djventilator | 9080.6 | 12003.6 | 9510.5 | 0.0 |

*Figure 4 Distance matrix*

### 4.2.2. Clustering

Clustering involves partitioning datapoints, into groups or clusters based on their similarity, such that datapoints within the same cluster are more similar to each other than to those in other clusters (Priy, n.d.). Clustering algorithms use mathematical techniques to determine optimal clusters based on similarity or dissimilarity between datapoints. In this case the datapoints on which the clustering will be applied are Twitter users. Similarity between users can be measured using various features, such as age, gender, and geographical location. This research focuses specifically on geographical location as a feature, from which the distance between Twitter users can be derived as a measure to assess their similarity. Low values for distance between Twitter users correspond with high values of similarity between users and vice versa. As a result, the obtained groupings from clustering are characterized by a significant number of Twitter users living in close spatial proximity to each other. In this study, these groups of Twitter users are also referred to as geolocated communities. The specific clustering algorithm used and details about the way it works are explained below.

#### 4.2.2.1. DBSCAN

Spatial clustering methods are used to group a set of spatial objects, such as coordinates, into clusters (Pattnaik, 2020). Among these methods, density-based clustering methods, including DBSCAN (Density-Based Spatial Clustering of Applications with Noise), are highly regarded for identifying geolocated communities on Twitter. The DBSCAN clustering algorithm is widely used for this task due to its advantages, such as the ability to identify clusters of different sizes and shapes. Additionally, in contrast to a lot of other clustering algorithms, DBSCAN doesn't require prior assumptions about the number of clusters (Pindado & Barrena, 2020).

#### 4.2.2.2. Algorithm

The DBSCAN clustering algorithm will also be used in this study to detect geolocated communities on Twitter. To describe the DBSCAN clustering algorithm, we need the following concepts based on a data set $D$ of Twitter users:

(1) The Twitter users within a radius $\varepsilon$ of a given Twitter user $u(u \in D)$ is the subset, denoted by $Neighborhood_{\varepsilon}(u)$, defined as: $Neighborhood_{\varepsilon}(u) = \{v \in D | dist(u, v) \leq \varepsilon\}$

(2) A Twitter user $u(u \in D)$ is denoted as a core Twitter user if $Neighborhood_{\varepsilon}(u)$ contains at least $MinUsrs$ Twitter users.

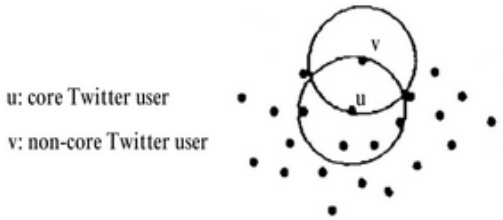


*Figure 5 Core and non-core Twitter users*

(3) A Twitter user $v(v \in D)$ is denoted as directly density-reachable from the Twitter user $u(u \in D)$ if $v$ is in $Neighborhood_{\varepsilon}(u)$ and $u$ is a core Twitter user.

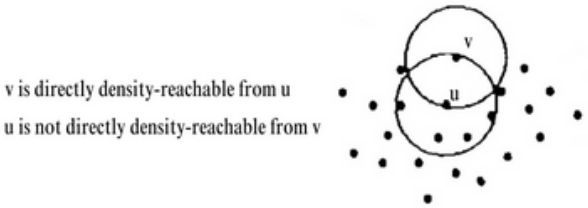


*Figure 6 Directly density-reachable Twitter users*

(4) A Twitter user $v(v \in D)$ is denoted as density-reachable from the Twitter user $u(u \in D)$ if there is a chain of Twitter users $u_1, \ldots, u_n$with $u_1 = u$ and $u_n = v$ such that $u_{i+1}$ is directly density-reachable from $u_i$.

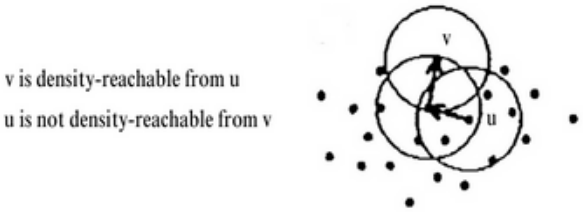


*Figure 7 Density-reachable Twitter users*

(5) Two Twitter users, $u(u \in D)$ and $v(v \in D)$ are denoted as density connected Twitter users, if there exists a Twitter user $w(w \in D)$ such that $u$ and $v$ are density-reachable from $w$.

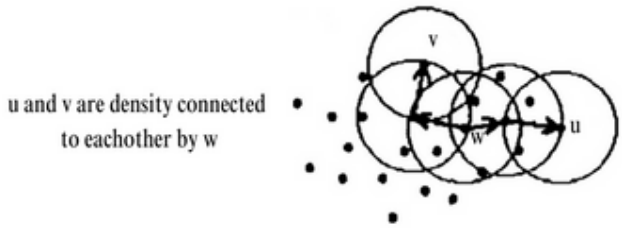


*Figure 8 Density connected Twitter users*

If there are any core Twitter users in the data set of Twitter users, DBSCAN starts by selecting a random core Twitter user $u(u \in D)$. DBSCAN finds all density-reachable Twitter users from $u$ and assigns them to the first cluster. If there are core Twitter users left that haven't been assigned to the first cluster, a new random core Twitter user $v(v \in D)$ that hasn't been assigned to the first cluster is selected. DBSCAN again finds all density-reachable Twitter users from $v$ and assigns them to the second cluster. If there are core points left that haven't been assigned to a cluster, a new random core Twitter user $w(w \in D)$ that hasn't been assigned to a cluster is selected. DBSCAN again finds all density-reachable Twitter users from $w$ and assigns them to the third cluster. The clustering procedure repeats itself until there are no core Twitter users left that are not assigned to any cluster.

In the end each core user is assigned to a cluster. Every non-core users that is density-reachable from at least one core user is also assigned to a cluster, and these non-core users are denoted as border Twitter users. The non-core users that are not density-reachable from any core Twitter users will not be assigned to any cluster. These Twitter users are too far away from the high density regions of Twitter users and are referred to as noise Twitter users. Figure 9 can help to clarify the algorithm as well as the concepts core, border and noise Twitter user (Ester, Kriegel, Sander & Xu, 1996; Pindado & Barrena, 2020).

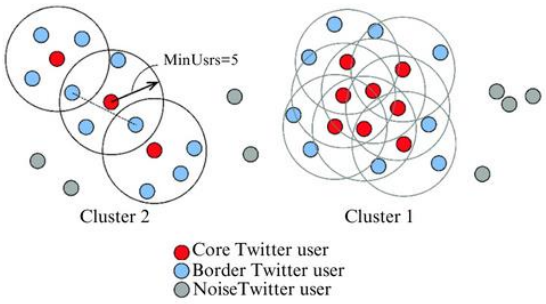


*Figure 9 Core, border and noise Twitter users*

Because every cluster starts from a core user, the minimal cluster exists of at least $MinUsrs$ with a maximal distance between them of $2\varepsilon$, but, clusters will often be larger than the minimal cluster. Starting from a core user, a clusters will expand by forming chains of users, absorbing every density-reachable user, much like a virus spreads itself over the map until there are no more people

in a close enough range to keep spreading. By understanding the DBSCAN clustering algorithm, it is clear that that resulting cluster of Twitter users are groups of Twitter users that live in a high density region. These high density regions are characterized by a significant amount of users that live in a relatively close spatial proximity to each other, in this study referred to as geolocated communities of Twitter users.

#### 4.2.2.3. Parameter selection

The DBSCAN algorithm requires the selection of two parameters in advance: $MinUsrs$ and $\varepsilon$. $MinUsrs$ represents the minimum number of Twitter users required in the neighbourhood of a given Twitter user for that user to be considered a core Twitter user. There are not many tools at hand for deciding upon the $MinUsrs$. Ultimately, the selection of the $MinUsrs$ value should rely on domain knowledge and familiarity with the dataset (Mullin, 2020). Considering that the dataset consists of 9415 Twitter users, a value of 20 for $MinUsrs$ seems reasonable. This determination was also made by iteratively adjusting the value and observing the resultant cluster distributions when visualized on a world map.

Given the value for $MinUsrs$, the value of $\varepsilon$ can be determined. $\varepsilon$ represents the maximum distance between two Twitter users for them to be considered within each other's neighbourhood. To select an appropriate value for $\varepsilon$ given a value for $MinUsrs$, the k-distance plot is often utilized (see Figure 10). A k-distance plot displays the distance to the k-th nearest neighbour for each user, plotted in ascending order. If $MinUsrs$ is chosen as the value for k and all distances to the k-th nearest neighbour are considered as potential values for $\varepsilon$, then the plot shows how many users will be identified as core and non-core users for each potential value of $\varepsilon$.

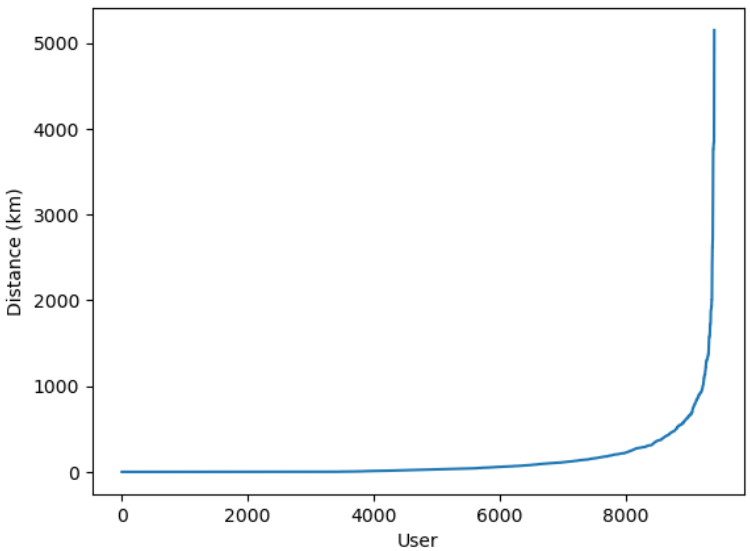


*Figure 10 k-distance plot (k=20)*

For a given user on the graph, if the distance to their k-th nearest neighbour is selected as $\varepsilon$, then that user would be considered a core user because they have k neighbours within a range of $\varepsilon$km. Since the graph is in ascending order, all users to the left of this user will also be core users. Users to the right of this user are non-core users, with those closest to the right having the potential to be classified as border users within a cluster, while those furthest to the right are likely to be classified as noise users and not included in any cluster (Ester et al., 1996; Pindado & Barrena, 2020).

It is important to strike a balance when selecting the value of $\varepsilon$. If too many Twitter users are classified as noise users, then there will be fewer users classified within clusters, resulting in less data available for research. On the other hand, if too many users are included in clusters, then some users may be too far away from other members of their cluster and not fit well within it.

A k-distance plot typically exhibits a sudden increase in distance to the k-th nearest neighbour (a "knee"). Users to the right of this knee have a much larger distance to their k-th nearest neighbour compared to the majority of Twitter users. The distance at the bottom of the knee is a good value for eps because it includes the majority of Twitter users while omitting outliers. In this example, a value of 500 was chosen for $\varepsilon$. From a total of 9415, Twitter users with known location, 9025 are classified in a cluster. Resulting in 390 noise Twitter users. Among the classified Twitter users, 8747 are core Twitter users, and 278 are border Twitter users.

## 4.3. Geolocated communities and their attitudes

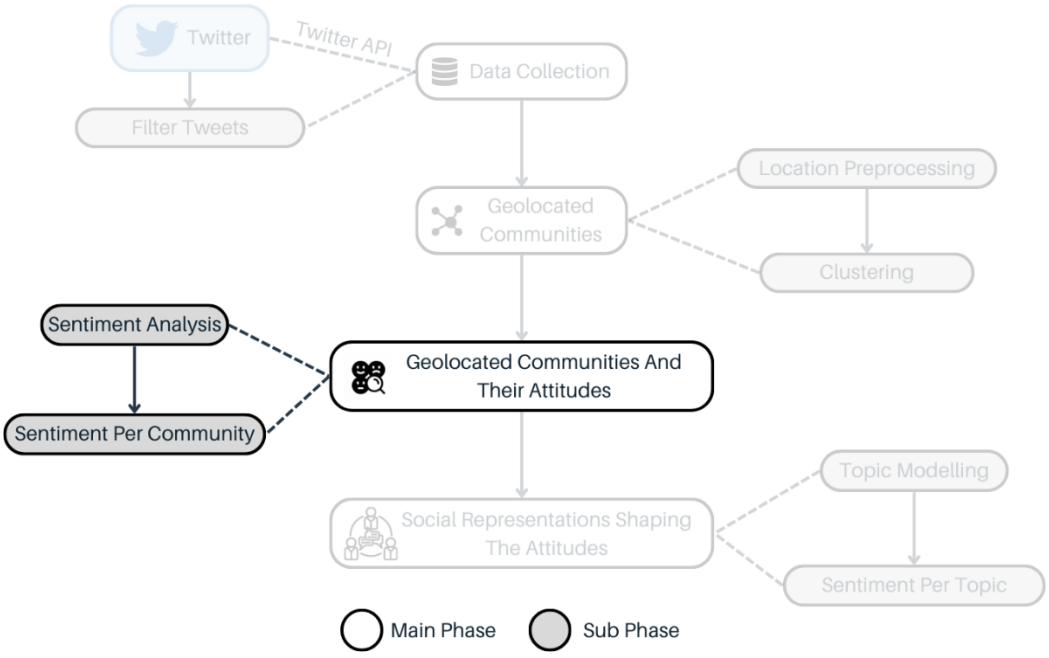


*Figure 11 Research methodology diagram: Geolocated Communities And Their Attitudes*

This subsection covers how the attitude of a tweet towards a certain food trend are derived with sentiment analysis. Afterwards there is explained how the overall attitude towards a food trend for each of the previously identified geolocated communities is derived.

### 4.3.1. Sentiment analysis

Sentiment analysis is a natural language processing (NLP) technique that involves using computational methods to understand the sentiment expressed in text. The goal is to determine whether the conveyed opinion is positive or negative. In the context of Tweets, sentiment analysis can provide insight into people's attitudes about a particular subject. There are two main approaches to performing sentiment analysis: the Lexicon-based approach and the machine learning-based approach. Each approach has its own strengths and weaknesses, and a trade-off must be made to determine which approach is most suitable for a specific use case. As the dataset of Tweets is not labelled with true sentiment values, it is not possible to calculate a metric that evaluates performance on this dataframe or compare different approaches in a quantitative way. Therefore, to make an informed decision about which sentiment analysis technique to use, a brief understanding of both methods is required.

Lexicon-based sentiment analysis is an unsupervised approach that relies on pre-defined sentiment lexicons. These lexicons contain sentiment scores or labels assigned to individual words. These scores are assigned on a numeric scale often ranging from negative numbers to positive numbers which can be associated with a semantic scale ranging from negative sentiment to positive sentiment respectively. As such, the numbers are an indication of the semantic meaning of the word. Aggregating the sentiment of a given Tweet can be accomplished by taking the average of the scores of its constituent words. Lexicon-based approaches can be computationally efficient, especially for real-time sentiment analysis. However, they may face challenges in handling sarcasm, context-dependent sentiments, humour or slang. (Bogaert, 2022)

In contrast to the lexicon-based approach, the machine learning approach to sentiment analysis is a form of supervised learning. Sentiment analysis machine learning models are trained on datasets of text items with their corresponding sentiment label (negative, neutral, positive). During training, the machine learning algorithm learns the relation between the text as an independent variable and its corresponding sentiment label as a dependent variable. This results in a classification model that outputs a sentiment label (negative, neutral or positive) corresponding with a given piece of text as input. However, the algorithm to train a machine learning model cannot perform computations with raw text data. Therefore, the raw text should be transformed by a feature extractor into a feature vector that captures the important information about the text in a numerical way. Once the model demonstrates satisfactory performance, it can be deployed to classify the sentiment of new, unseen

text inputs (expressed as feature vectors). Figure 12 can help understand the process of making such a model (Bogaert, 2022).

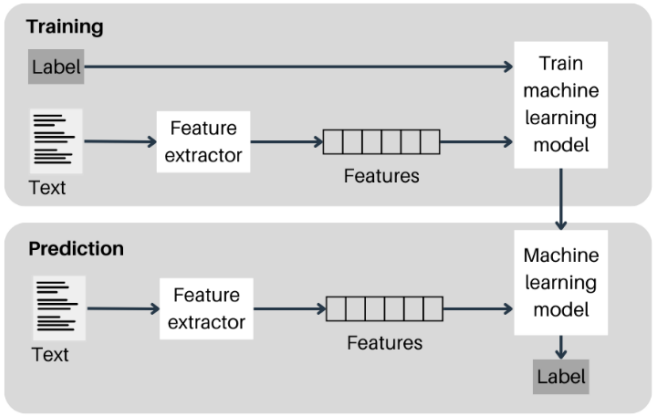


*Figure 12 Sentiment analysis: machine learning approach*

Machine learning approaches for sentiment analysis are capable of addressing the limitations of lexicon-based approaches, such as handling sarcasm, context-dependent sentiments, humour and slang. Machine learning models are generally more accurate when trained for specific applications. However, training a machine learning model requires significantly more time and resources than a lexicon-based approach, as it necessitates more computational power and the acquisition of labelled datasets can be costly. Nonetheless, the availability of numerous pre-trained models that are freely usable renders model accuracy the decisive factor in favour of machine learning models for sentiment analysis. In the following section, we provide an overview of the specific model employed. It is worth noting that a comprehensive elaboration of a lexicon-based approach is also included in the code on GitHub (Alternatives.ipynb). However, since these sentiment values are not utilized for further analysis, we do not provide an exhaustive explanation here.

#### 4.3.1.1. Twitter-XLM-roBERTa-base sentiment model

When selecting a pre-trained machine learning model for sentiment analysis, it is crucial to choose a model that has been trained on domain-specific data that is similar to the data on which the model will be applied. For instance, a model that has been trained on a labelled dataset of French product reviews may not be very accurate in determining the sentiment of Tweets written in multiple languages. In our case, we aim to determine the sentiment of Tweets written in multiple languages. The Twitter-XLM-roBERTa-base sentiment model is a pre-trained machine learning model for sentiment analysis developed by Cardiff NLP. It is a model that has been trained to detect the sentiment of Tweets in multiple languages, making it suitable for use in our case (Barbieri et al., 2022).

The Twitter-XLM-roBERTa-base sentiment model differs from traditional machine learning models for sentiment analysis. The model was built in a similar fashion as the process shown in figure 12, however now the machine learning model isn't built as a separate model, instead is built on top of the feature extractor. The feature extractor and the machine learning model together form one big machine learning model that takes raw text as input and produces a sentiment label as output. To understand how it is build, an understanding of the concepts language model, and fine-tuning is needed. In the following section, we provide an explanation of how it was built.

#### 4.3.1.2. Language model for text embeddings

The reason for introducing language models is that they are used as feature extractors (figure 12). Although there are various other techniques that can be used to transform text into feature vectors such as bag-of-words, tf-idf, and doc2vec, language models are the state-of-the-art methods for extracting feature vectors from text that capture most important information about the text such as context, meaning, sentiment, and subject in a numerical way.

A language model is a type of neural network that comprehends human language and can serve various purposes. One such language model is XLM-roBERTa, which belongs to the category of transformer language models. XLM-roBERTa was pre-trained on a large corpus of text using the Masked Language Modeling (MLM) objective. The model takes text as input and randomly masks 15% of the words. The partly masked input text is then processed by the neural network to predict the masked words. Since the model masks the words itself, it knows the true output and can create input-output pairs automatically. This is known as self-supervised training and allows the model to be trained on raw text without the need of manual labelling by humans. XLM-roBERTa was trained on 2.5 TB of textual data across 100 languages. The resulting XLM-roBERTa language model can predict masked words in a given input text by assigning a probability to each word in its vocabulary that it is the masked word. The vocabulary consists of all the words known by the model based on the training data it has seen (Barbieri et al., 2022).

A language model comprehends human language based on the textual data on which it has been trained. However, linguistic patterns, sentence construction, and vocabulary usage vary across different domains. For instance, specific hashtags such as #BlackLivesMatter may frequently appear in a large corpus of Tweets but are less likely to be found in a corpus of Wikipedia text. The training dataset for XLM-roBERTa comprised Wikipedia text, Reddit posts, books, newspapers, and other sources. Despite the utilization of extensive training data from various sources, social media data - specifically Twitter (the platform emphasized in this paper) - was excluded from the massive multilingual training. By continuing to pretrain the neural network from available checkpoints on new domain-specific data that the model has not previously encountered, a

language model can be adapted to its specific domain of employment. This can significantly enhance the model's performance for predicting the masked words of domain-specific input text. To adapt XLM-roBERTa and expand its knowledge of human language on Twitter, 198 million Tweets in different languages posted between May 2018 and March 2020 were used for further pretraining. The resulting model is referred to as Twitter-XLM-roBERTa (Barbieri et al., 2022).

Although the Twitter-XLM-roBERTa language model is capable of predicting masked words within a given text, it is mostly intended to extract features from non-masked text. During the training process, the neural network learns an internal representation of 100 languages within its hidden layers, which can be utilized for feature extraction. To obtain a feature vector representation of a given input text, the values of neurons within a specific layer corresponding to that input can be utilized. Figure 13 can help to clarify this concept but it is a simplified version to represent the network architecture of such a language model and does not entirely represent reality. However the model wasn't specifically trained with the objective of making good feature vectors, somehow the values of the neurons in a certain layer seem to capture a lot of information about the text. It may be necessary to consider the feature vector, also known as text embeddings, as a black box and accept that it captures information about the text without exactly understanding why. The Twitter-XLM-roBERTa language model serves as an efficacious feature extractor, capable of converting Tweets in various languages into feature vectors. Despite the fact that the resulting vectors are often large and difficult for human interpretation, they can be utilized for a range of downstream natural language processing (NLP) tasks, such as sentiment analysis, which will be discussed in the subsequent section (Barbieri et al., 2022, Zvornicanin, 2023).

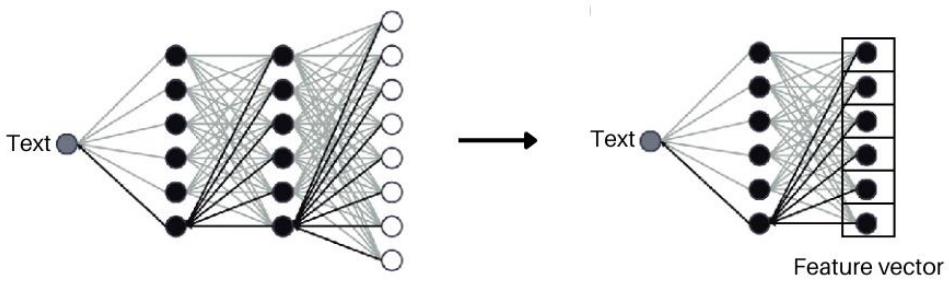


*Figure 13 Feature extraction from simplified neural network architecture of language model*

#### 4.3.1.3. Fine tuning for sentiment analysis

As previously mentioned, machine learning model for sentiment analysis can be trained in a supervised fashion on a sentiment-labelled dataset of Tweets represented as feature vectors (figure 12). Various machine learning techniques are available for learning the relationship between the feature vectors as independent variables and their corresponding sentiment labels as dependent

variables. These techniques include Naive Bayes, Logistic Regression, Support Vector Machines, Random Forest, XGBoost, and Artificial Neural Networks (Bogaert, 2022).

If an artificial neural network is chosen for the task, it can be built on top of the Twitter-XLM-roBERTa neural network (figure 14). For certain input text, Twitter-XLM-roBERTa outputs the feature vector as the values of neurons in a specific layer corresponding to that input text. Given that this feature vector or neurons of a certain layer will serve as the input to the other neural network used for sentiment classification, it is not necessary to build this sentiment classifier as a separate model. Instead, an additional layer can be added to the Twitter-XLM-roBERTa neural network (figure 14). The training of the parameters of this additional layer for sentiment analysis happens by freezing the layers of the first part of the network such that only the additional layer is adapted. This process essentially boils down to finetuning the Twitter-XLM-roBERTa model to perform sentiment analysis resulting in the Twitter-XLM-roBERTa-base sentiment model (Barbieri et al., 2022).

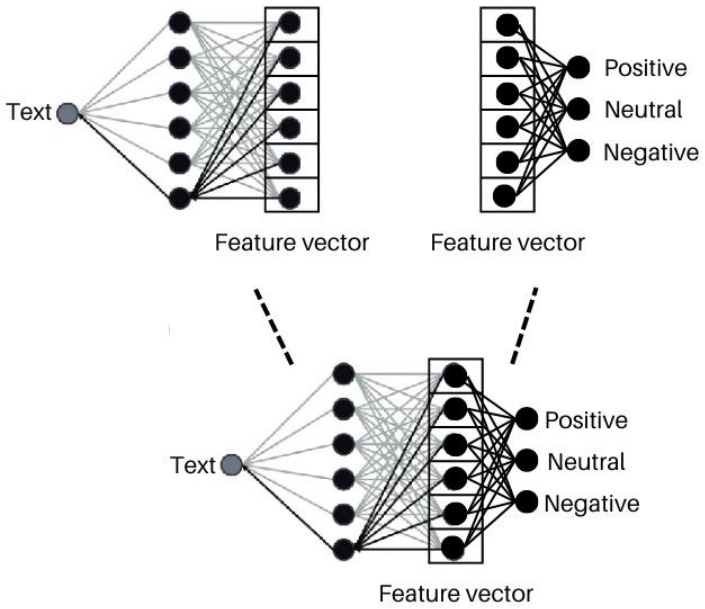


*Figure 14 Fine-tuning language models for sentiment analysis*

The practice of adapting a language model to perform a specific NLP task is commonly referred to as fine-tuning. Transformer language models trained using the masked language modelling objective, where the masked word can appear anywhere in the text (bidirectional), include BERT, roBERTa, XLM-roBERTa, Twitter-XLM-roBERTa, and others. These pre-trained models exhibit exceptional performance when fine-tuned for text classification tasks such as sentiment analysis, topic modelling, and named entity recognition. Conversely, transformer models trained with the objective of predicting the next word in a sequence of text where the masked word appears at the

end of the text (unidirectional) include GPT-2, GPT-3, and others. These pre-trained models are better suited for tasks such as text generation or translation (Statie, 2022).

Table 5 displays an example of the results obtained by assigning a sentiment score, ranging between -1 and 1, to each Tweet in the dataframe using the pretrained Twitter-XLM-roBERTa sentiment model. The sentiment score is calculated by subtracting the assigned probability of the Tweet being negative from the assigned probability of the Tweet being positive while neglecting the probability of the Tweet being neutral. This means that a negative score indicates negative sentiment, while a positive score indicates positive sentiment. As the dataset of Tweets was not labelled with ground truth, it is not possible to compute a metric to evaluate the performance. However, based on human judgment, an examination of the sentiment scores assigned to each Tweet suggests that the model performs well.

| text | sentiment |
|---|---|
| @MEGroenstege Gisteren bij #kassa. Diverse #vegan burgers getest. Eindconclusie: rubber en karton smaak. Het was de saus en broodje die het net eetbaar maakte. 1 Mac burger kreeg een voldoende, de rest een dikke onvoldoende. Noem het geen vlees, maar groente of #soja burgers. Nep vlees is het. | -0.245814 |
| Vegans who are animal rights activists are complete NPCs. #animalrightsactivists #vegan #animalrights #npcs #npc2023 | -0.275710 |
| 🥑Millions swear by going vegan as a remedy for skin problems, low energy and so much more. 🥗 Trying to jump on the plant based train? 🚆 Here's some helpful advice from those who have already succeeded! https://t.co/mYLQ90LkBn #PlantBased #Vegan #HealthyLiving https://t.co/RF85XjyuyR | 0.522947 |
| Ik ben niet #vegan, omnivoor of carnivoor, geen vegetariër of flexitariër. Labels suggereren dat ik geen keuze heb, dat ik niet anders kan. Ik kies er voor om #duurzaam te leven, zo eet ik meestal plantaardige eiwitten. Dit doe ik bewust en in vrijheid. https://t.co/FjQBdd5moW | 0.139362 |

*Table 5 Preview of sentiment analysis on Tweets*

### 4.3.2. Sentiment per community

After performing sentiment analysis on the dataset of Tweets related to the #vegan food trend, each Tweet in the dataset was associated with a sentiment score that represented the attitudes towards the food trend. Additionally, each Tweet was associated with a cluster that represented the geolocated community to which the Tweet author belongs. To investigate the distribution of attitudes towards the #vegan food trend across the world, we aggregated the data by geolocated community and calculated the average sentiment score for each community. This allowed us to examine how attitudes towards the #vegan food trend varies across different regions.

## 4.4. Social representations shaping the attitudes

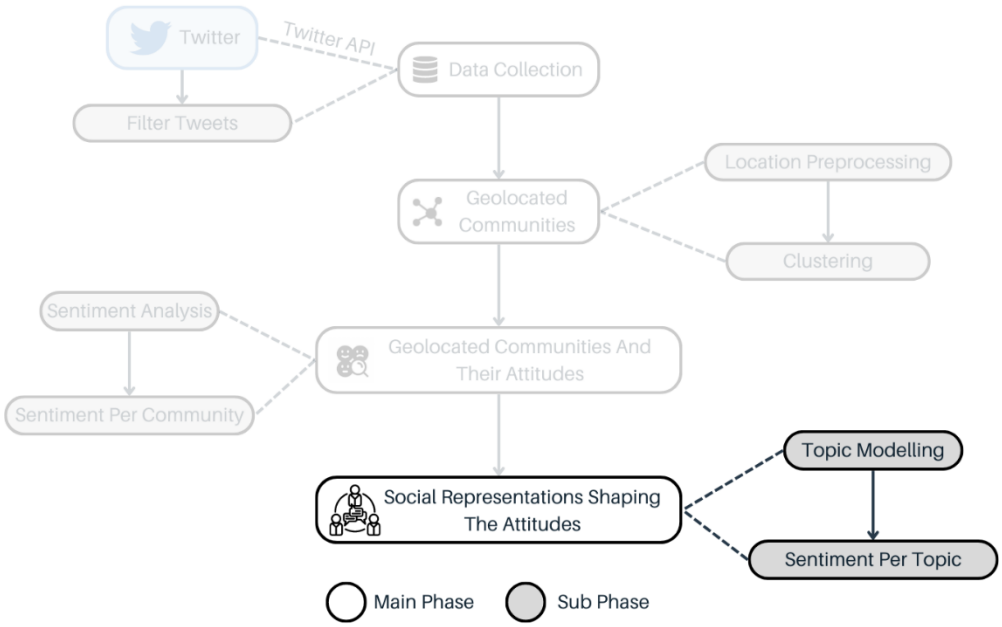


*Figure 15 Research methodology diagram: Social Representations Shaping The Attitudes*

This subsection covers how the existing social representations about certain food trend are discovered using topic modelling. Afterwards there is zoomed into how the tweets are aggregated to detect differences in attitude towards a food trend depending on the social representation.

### 4.4.1. Topic modelling

Topic modelling is a natural language processing (NLP) technique that involves using computational methods to discover the topics that occur in a collection of text, without knowing those topics at first. In the context of the #vegan trend on Twitter, topic modelling can provide insight into the main underlying themes and subjects being discussed from the chaos of Tweets. The discovered topics can be considered as subtopics, as veganism itself can also be seen as a topic. Various techniques exist for performing topic modelling, each with its own strengths and weaknesses. Prior to applying topic modelling, the number of topics and the specific topics to look for are not known. This means that the ground truth is not known for the dataframe of Tweets, making it difficult to calculate a metric that evaluates the performance of the various techniques. While some performance metrics exist to evaluate the performance of topic modelling, they often vary among techniques and cannot be used to compare them. Instead, they are more commonly used to fine-tune hyperparameters of a certain topic modelling technique. Often to assess the performance of a topic modelling technique, a qualitative approach is used where human interpretation can help to assess whether the identified topics are understandable and coherent. Therefore, to make an informed decision about which topic modelling technique to use, a brief understanding of some of the available techniques is required. This study considers Latent Dirichlet

Allocation (LDA), Gibbs Sampling Dirichlet Mixture Model (GSDMM), and clustering text embeddings, to decide which one is most suitable for grouping Tweets according to topic.

Latent Dirichlet Allocation (LDA) is a popular method for discovering topics in a corpus of text documents. It is based on the assumption that each document is a mixture of topics, and each topic is a distribution of words. One of the main advantages of LDA is that it is an unsupervised learning technique, meaning that labels or categories for documents are not required. LDA can automatically infer the topics from the data and assign each document a probability of belonging to each topic. This can reveal hidden patterns and insights that might not have been noticed otherwise. However, LDA also has limitations that should be considered before using it. First, it requires good pre-processing to work effectively, which can take a lot of effort. The goal is to ensure that the documents mostly consist of words that are potentially relevant to a topic. Steps to obtain this include removing stop words and punctuation, tokenization, term filtering, and lemmatization as well as other steps. These steps are essential to improve the quality and consistency of the data. Another disadvantage of LDA is that the number of topics must be given as input to the model. However, the number of topics in a corpus of text is often not known in advance. Further, LDA relies on the assumption that the words in each topic are related and meaningful, but this may not always be the case in reality. As a result, it can produce ambiguous or incoherent topics. Finally, despite its great results on medium to large sized texts like news articles or emails, LDA performs poorly on short documents like Tweets. This is because LDA assumes multiple topics per document; however, short texts often only cover one topic (Bogaert, 2022).

Gibbs Sampling Dirichlet Mixture Model (GSDMM) is an altered version of LDA that assumes each text document consists of only one topic instead of multiple topics. In this way, GSDMM can compensate for the shortcomings of LDA for short text topic modelling (STTM) (Amrouche, 2019).

Another approach to topic modelling involves clustering together text embeddings made with language models (Grootendorst, 2020). This approach consists of three steps: generating the embeddings, clustering them together, and finally retrieving the most relevant words that belong to the topics. This approach does not require a lot of text pre-processing because language models that are pretrained on domain-specific text data can make very good embeddings from raw text that capture the semantics. Additionally, there is no need to make prior assumptions about the number of topics because various clustering methods do not need to know the number of clusters to detect, such as DBSCAN. Furthermore, not every text document in the corpus of text is forced to belong to a cluster; as previously seen, clustering methods like DBSCAN do not cluster noise points. Finally, language models can work very well on small texts. While building a language model requires

large computation power, many pretrained language models are available for use. This means that this approach addresses all the disadvantages of both previous methods and will be continued with for topic modelling in this study. The three steps are explained in more detail in the following paragraphs. Note that LDA has been worked out in the code on GitHub (Alternatives.ipynb); however, since these results are not continued with, it is not explained in further detail.

#### 4.4.1.1. Generating text embeddings

As previously mentioned, text embeddings are high-dimensional feature vectors that capture a lot of information about the text in a numerical way. Each dimension of the embedding captures a different aspect of the text's semantics. Since each dimension of the embedding captures a different aspect of the text's semantics, embeddings that are close to each other in vector space are likely to be similar in semantics. The same can be done with single words, where each word is represented as a vector that captures the semantics of that word. Figure 16 shows an example of such word embeddings and how the dimensions can be given meaning. From this understanding about text embeddings, it is intuitively clear that texts about similar topics will have embeddings that are close to each other in the high-dimensional space.

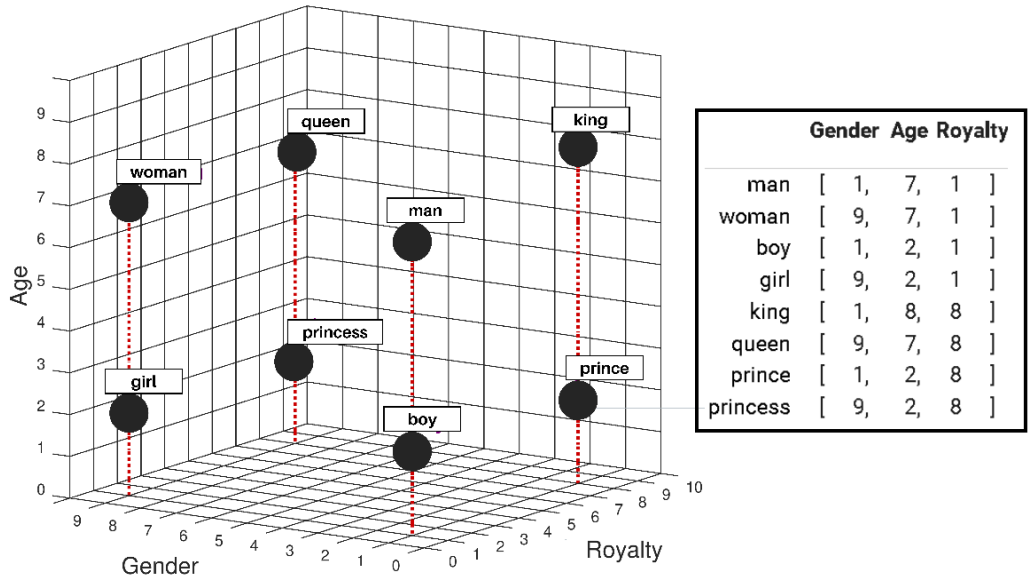


*Figure 16 Word embeddings in 3-dimensional vector space*

It has already been discussed how language models can be used for extracting text embeddings from text. The language model used for retrieving text embeddings from Tweets is called paraphrase-multilingual-MiniLM-L12-v2. A model like the previously seen Twitter-XLM-roBERTa could also be used to achieve decent performance. Paraphrase-multilingual-MiniLM-L12-v2 is a similar language model trained for the masked language modelling objective but finetuned on the specific task of making text embeddings in a way that similar texts have

embeddings that are close to each other in vector space. If the pretrained model were not to be finetuned, it would already make the embeddings in a way that similar texts have embeddings that are closer to each other, but finetuning for this specific task can make this even more prevalent (Reimers & Gurevych, 2020).

The fine-tuning of the model happens in a supervised way where a labelled dataset is needed with text pairs and a label that indicates their true similarity. A value of 1 indicates that the vectors are identical, while a value of -1 indicates that they are completely dissimilar. The pretrained network can then be fine-tuned with a Siamese Network Architecture (figure 17). For each text pair, text A and text B are passed through the network, yielding embeddings u and v. The similarity of these embeddings is computed using cosine similarity and the result is compared to the true similarity label. This allows the network to be fine-tuned and make embeddings close to each other in terms of cosine similarity for texts that are similar and vice versa (Reimers & Gurevych, 2020).

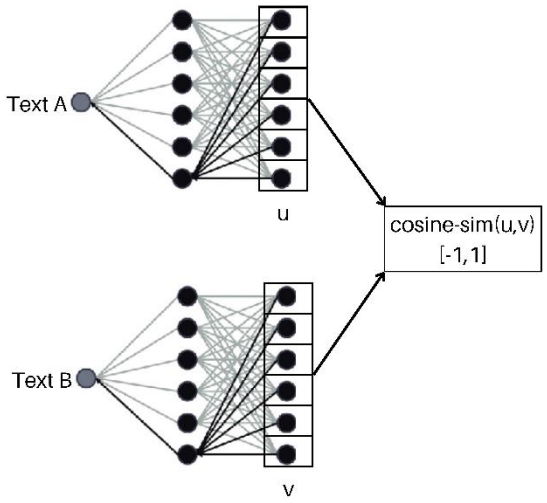


*Figure 17 Fine-tuning of language model for better word embeddings*

It is also important to mention that UMAP was used to reduce the dimensions of the obtained text embeddings because many clustering algorithms handle high dimensionality poorly. The dimensions were reduced from 384 to 5.

#### 4.4.1.2. Clustering embeddings

Texts whose embeddings are close to each other in the 5D vector space can be grouped together by making use of a spatial clustering algorithm. The resulting groups are considered as texts about the same topic. The clustering algorithm used is Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), an adapted version of DBSCAN. The difference with DBSCAN is that HDBSCAN only takes the minimum cluster size as an input parameter, where epsilon can vary, which makes it able to identify clusters with varying density. Like DBSCAN,

HDBSCAN does not force every Tweet to belong to a cluster, so some Tweets will be seen as noise not fitting into any cluster. After applying HDBSCAN, groups of Tweets are obtained.

#### 4.4.1.3. Most important words

The goal of topic modelling might be to group Tweets according to their topic, as done in precious section, or to discover what the discussed topics are about. This study is especially interested in the discovery of discussed topics. To discover the subtopics people talk about within the topic of veganism, it is necessary to uncover the most important words that distinguish one cluster of Tweets from others. To do this, c-TF-IDF can be used, a class-based variant of TF-IDF. When TF-IDF is applied as usual on a set of documents, it compares the importance of words between documents. Instead, all documents in a single topic group can be treated as a single document and then TF-IDF can be applied. The result would be a very long document per topic group since all text is added together. The resulting TF-IDF score would demonstrate the important words in a topic.

### 4.4.2. Sentiment per topics

Each tweet in the dataset now has a topic label based on the topic the tweets was assigned to. The content of the topics is also known because of c-TF-IDF. In the part about sentiment analysis there is discussed how each Tweet was associated with a sentiment score that represented positive or negative attitudes. To investigate the distribution of sentiment across topics within the #vegan food trend, the data was aggregated by topic and the average sentiment score was calculated for each topic. This allowed for an examination of how topics within the #vegan food trend are associated with positive or negative sentiment.

# 5. Results

After running through the methodology, it is now time to examine the obtained results, which will be presented in the order in which the methodology was performed.

## 5.1. Geolocated communities

Figure 18 shows all the Twitter users that have been assigned to a cluster plotted on the world map. In this plot, the noise Twitter users were omitted since they have not been assigned to a cluster and appear to be all over the world in non-dense regions. It can be noticed that there are significant differences with respect to the sizes of the clusters. North America and Europe both contain many more Twitter users compared to other regions. From this plot, it is not possible to determine how much people engage with the veganism food trend around the world. It would be biased since the

areas with the most Twitter activity in general will largely determine where the clusters are situated, regardless of the trend in consideration.

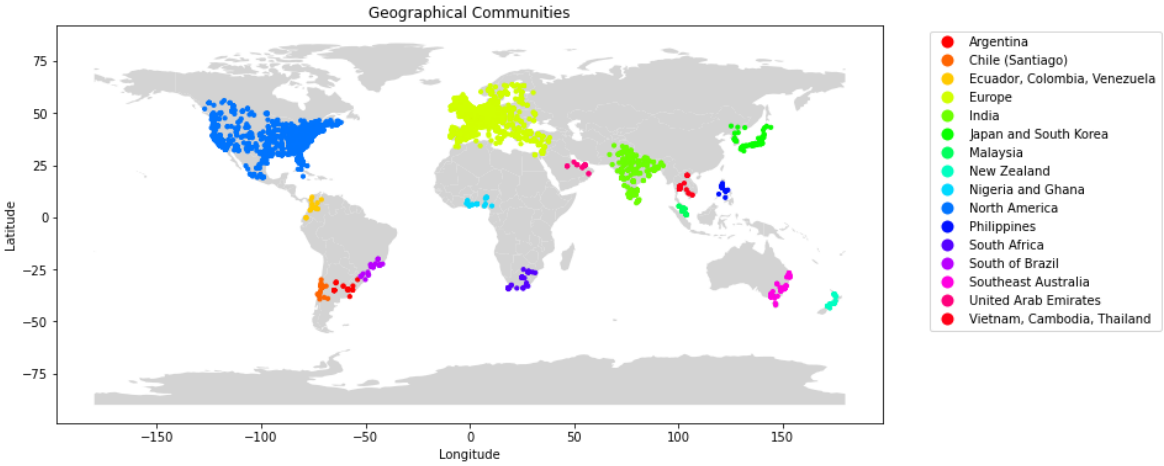


*Figure 18 Geolocated communities on Twitter*

Although the plot in figure 18 does not allow for conclusions about where people engage the most with veganism, it does allow for the distinction of different areas where a significant number of people live together in relatively close spatial proximity to each other. These detected regions largely fall within the borders of different socio-cultural regions if we go by figure 19 (Bell, 2013).


*Figure 19 World cultural regions*

## 5.2. Geolocated communities and their attitudes

This makes it interesting to discover how the obtained sentiment scores are distributed across the different regions and to discover differences and similarities. This is represented in table 6. It can be observed that the regions that fall within the same cultural region as depicted in figure 19 seem to have similar sentiment scores for their Tweets about veganism. It is important to keep in mind

that due to the low number of Twitter users in some clusters, the results are difficult to generalize over the whole community.

| Cluster Size | Region | Sentiment (Mean) | Sentiment (Std) |
|---|---|---|---|
| 83 | United Arab Emirates | 0.410891 | 0.292365 |
| 665 | Japan and South Korea | 0.406882 | 0.372077 |
| 136 | South Africa | 0.262754 | 0.413241 |
| 78 | Philippines | 0.262558 | 0.377367 |
| 50 | Nigeria and Ghana | 0.256810 | 0.311818 |
| 13572 | North America | 0.244034 | 0.414379 |
| 14002 | Europe | 0.238360 | 0.470702 |
| 109 | South of Brazil | 0.234407 | 0.419127 |
| 486 | Vietnam, Cambodia, Thailand | 0.217437 | 0.351495 |
| 1134 | Noise | 0.183505 | 0.457608 |
| 315 | Southeast Australia | 0.183192 | 0.547036 |
| 51 | Ecuador, Colombia, Venezuela | 0.169843 | 0.540855 |
| 1545 | India | 0.156795 | 0.417460 |
| 437 | Argentina | 0.130484 | 0.240458 |
| 57 | Chile (Santiago) | 0.121637 | 0.468446 |
| 121 | Malaysia | 0.076633 | 0.349719 |
| 172 | New Zealand | 0.046083 | 0.634782 |

*Table 6 Sentiment per community*

## 5.3. Social representations shaping attitudes

After knowing the sentiment scores per region, it is interesting to discover what the main topics are that are discussed within Tweets about veganism and to see how sentiment scores vary across topics. This is represented in table 7. First of all, very good and coherent topics are observed. From human interpretation of the topic by looking at the most important words per topic, several coherent topics can be clearly distinguished. If the results are evaluated with human interpretation, it can be said that the topic modelling technique performs very well. On the other hand, a lot of noise is noticed, meaning that many Tweets are not assigned to a topic. To make very coherent clusters, there had to be little tolerance for including noise points. When assessing the sentiment scores per topic, it is noticed that some sentiment scores differ a lot across topics. So some topics are talked about with a negative connotation, whereas others have a positive connotation.

| Topic Size | Most Frequent Words | Sentiment (Mean) | Sentiment (Std) |
|---|---|---|---|
| 280 | lunch, dinner, weekend, saturday, open | 0.514285 | 0.359587 |
| 1363 | chocolate, cake, cookies, coconut, recipe | 0.435970 | 0.296184 |
| 256 | pasta, noodles, spaghetti, recipe, sauce | 0.401892 | 0.266049 |
| 479 | salad, recipe, dressing, salads, red | 0.375659 | 0.295136 |
| 394 | tofu, rice, sauce, recipe, stir | 0.374212 | 0.287870 |
| 267 | pizza, veganpizza, pizzas, vegan, pepperoni | 0.366810 | 0.367587 |
| 217 | cheese, cheeses, based, plantbased, plant | 0.366799 | 0.405563 |
| 376 | burger, burgers, sandwich, patty, fries | 0.355682 | 0.378229 |
| 1024 | health, healthy, nutrition, diet, healthyfood | 0.277909 | 0.308772 |
| 22266 | noise | 0.260272 | 0.408983 |
| 298 | protein, plant, sources, based, powder | 0.237020 | 0.409532 |
| 2476 | vegan, veganism, vegans, food, govegan | 0.170044 | 0.516951 |
| 249 | soupbase, vegetablebroth, seafoodbroth, redibasecooking, lowsodium | 0.134403 | 0.086143 |
| 1328 | govegan, vegan, veganfortheanimals, animaliliberi, heute | 0.097326 | 0.551070 |
| 312 | milk, dairy, cows, dairycows, dairyfarm | 0.067311 | 0.505165 |
| 1428 | animals, animal, meat, govegan, animalrights | -0.337929 | 0.529524 |

*Table 7 Sentiment per topic*

The Tweets were clustered in a 5-dimensional space, but the Tweet feature vectors can be reduced to 2 dimensions, allowing for a graphical overview (figure 20). It is important to keep in mind that this 2-dimensional space was not used for the clustering as such; however, looking at the graph, it seems like this could also have yielded some good results. Every point in figure 20 represents a Tweet by plotting its feature vector with only two numerical features in a 2-dimensional frame. The axes are not shown since it is difficult to determine what they mean and for this reason they are irrelevant, but one can still hypothesize about their meaning by using human interpretation. This is a nice graphical way to represent the topics in a lower-dimensional space.

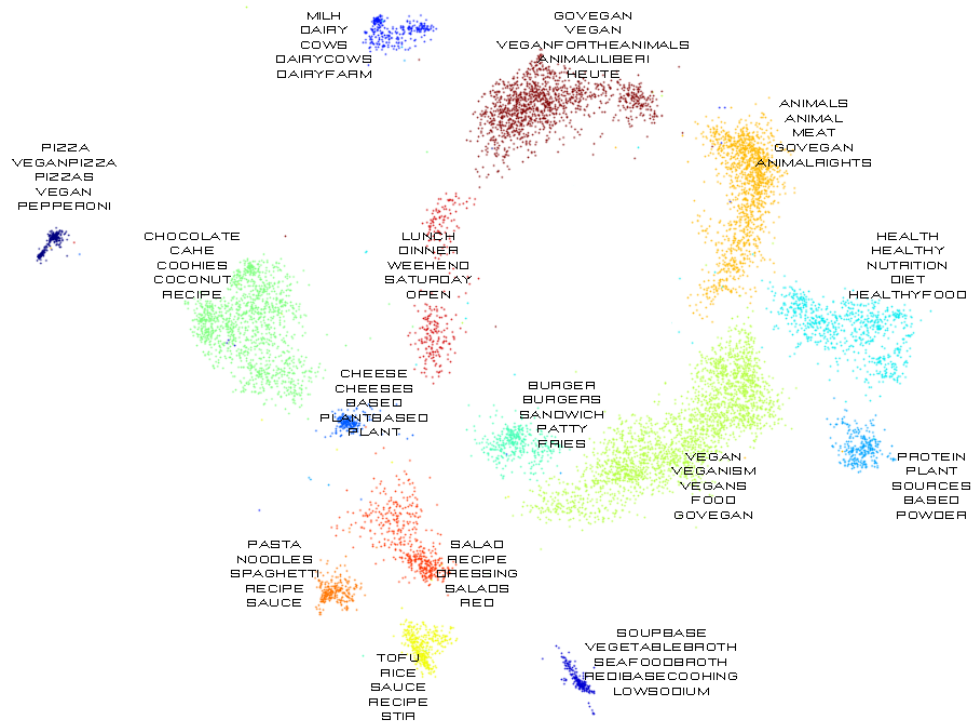


*Figure 20 Visualization of topic clusters in 2 dimensions*

# 6. Limitations

As previously highlighted, social representations and the consequent attitudes towards a specific subject are formed through social interactions within a community. However, this study only focuses on a significant number of users who live in close spatial proximity to each other as a criterion for community detection, not taking into account the possibility that individuals who live in the same neighbourhood may not necessarily interact with one another. The idea of communities as groups of people who live in the same neighbourhood, know and support each other is a nostalgic and romanticized notion from the past. In fact, most people do not know or interact with their neighbours, and their social ties often extend far beyond their immediate geographic location (Gruzd, Jacobson, Wellman, & Mai, 2016). The perception of community has been expanded and redefined over time, with politicians using the term to refer to groups of people with similar characteristics or identities, and social scientists conceptualizing communities as imagined or mental constructs based on shared communication and identity (Gruzd et al., 2016).

Because the DBSCAN clustering algorithm only considers a geographical dimension to capture communities, the identified geolocated communities of Twitter users might not be entirely in line with the current definition of a community of people. It is impossible to set strict, unambiguous boundaries of geolocated regions where people within that region are part of the same community, as communities are now understood to be much more fluid and diverse than previously thought.

Yet still, in the context of food trends, geographical location is an interesting element to take into consideration since food traditions vary greatly based on location. For example, Italy is known for pizza, Mexico for tacos, and Unites States for burgers.

Another limitation is caused by the fact that users can manually enter their location on Twitter. First, some users may not enter their location at all since it is not obligatory to do so on Twitter. Additionally, users may not enter their real location or they may spell it incorrectly such that the program cannot detect which location they are referring to. Finally, the geographical distribution of Tweets about a certain topic may be biased because only Tweets based on English hashtags are retrieved, causing most of the retrieved Tweets to originate from English-speaking countries. For example, #vegan would be #vegano or #vegana in Spanish. The study could be improved by taking this into consideration when retrieving tweets.

A final limitation that is worth mentioning relates to the topic modelling. In the results from the topic modelling, some very coherent topics had been noticed. However, this came at the price of omitting a lot of tweets from belonging to any topic. This also gives us fewer tweets per topic to take the average sentiment of, which is why the results might be less reliable.

# 7. Conclusion

In conclusion, this thesis has demonstrated the value of some state-of-the-art techniques to perform sentiment analysis and topic modelling. The study was able to very accurately determine the sentiment of Tweets in multiple languages where the model was able to handle sarcasm, context-dependent sentiments, humour and slang effortlessly. With regard to topic modelling some very coherent subtopics withing the #vegan food trend on Twitter had been distinguished. Both methods made use of a language model that was finetuned to perform the task of interest. The advent of language models has had a profound impact on the field of natural language processing. While the efficacy of one model may surpass that of another for a particular task, the overall influence of language models on NLP tasks has been significant. These models have consistently demonstrated superior performance for almost any NLP tasks in comparison to methods employed prior to their introduction (Bush, 2022). DBSCAN has also shown to perform well for detecting geolocated communities on Twitter despite some of the limitations in the dataset as mentioned in section 6.

The adopted approach towards detecting geolocated communities, sentiment analysis, and topic modelling have shown their strengths as separate techniques. To enrich the study and add some sort of storyline through the research, the tree methods used where combined to see if relations can be discovered between the geolocated communities, their sentiment and if sentiment depends on the topics discussed. With regard to this integrated approach some interesting results were obtained as

sentiment seems to vary between certain geolocated communities, and the sentiment conveyed in a Tweets seems to depend on the topic it addresses. While the findings are limited by the available data, they provide a promising foundation for future research.

# References

Amrouche, M. (2019, August 22). Short Text Topic Modeling. Towards Data Science. https://towardsdatascience.com/short-text-topic-modeling-70e50a57c883

Bäckström, A., Pirttilä-Backman, A.-M., & Tuorila, H. (2004). Willingness to try new foods as predicted by social representations and attitude and trait scales. Appetite, 43(1), 75-83. https://doi.org/10.1016/j.appet.2004.03.004

Bakillah, M., Li, R.-Y., & Liang, S. H. L. (2015). Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: The case study of Typhoon Haiyan. International Journal of Geographical Information Science, 29(2), 258-279. https://doi.org/10.1080/13658816.2014.964247

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 258-266). European Language Resources Association. https://aclanthology.org/2022.lrec-1.27

Bartels, J., & Reinders, M. J. (2010). Social identification, social representations, and consumer innovativeness in an organic food context: A cross-national comparison. Food Quality and Preference, 21(4), 347-352. https://doi.org/10.1016/j.foodqual.2009.08.016.

Bell, J. M. (2013). Cultural Regions. In K. L. Lerner, B. W. Lerner, & S. Benson (Eds.), Human Geography: People and the Environment (Vol. 1, pp. 190-193). Gale. https://link.gale.com/apps/doc/CX2062300074/GVRL?u=oreg77062&sid=bookmark-GVRL&xid=043724ac

Bogaert, M. (2022). Lecture 4 [PowerPoint slides]. SOCIAL MEDIA AND WEB ANALYTICS. (Unpublished lecture slides)

Bogaert, M. (2022). Lecture 5 [PowerPoint slides]. SOCIAL MEDIA AND WEB ANALYTICS. (Unpublished lecture slides)

Bush, N. (2022, August 29). Which NLP Task Does NOT Benefit From Pre-trained Language Models? Towards AI. https://towardsai.net/p/nlp/which-nlp-task-does-not-benefit-from-pre-trained-language-models

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. Computers, Environment and Urban Systems, 53, 47-64. https://doi.org/10.1016/j.compenvurbsys.2014.11.002

Deitrick, W., & Hu, W. (2013). Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks. Journal of Data Analysis and Information Processing, 01, 19-29. http://dx.doi.org/10.4236/jdaip.2013.13004

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: AAAI Press.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., Ju, Y., Gong, L., Adams, B., & Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science, 31(6), 1245-1271. https://doi.org/10.1080/13658816.2016.1273357

Grootendorst, M. (2020, October 5). Topic Modeling with BERT. Towards Data Science. https://towardsdatascience.com/topic-modeling-with-bert-779f7db187e6

Gruzd, A., Jacobson, J., Wellman, B., & Mai, P. (2016). Understanding communities in an age of social media: The good, the bad, and the complicated. Information, Communication & Society, 19(9), 1187-1193. https://doi.org/10.1080/1369118X.2016.1187195

Hridoy, S.A.A., Ekram, M.T., Islam, M.S., Ahmed, F., & Rahman, R.M. (2015). Localized Twitter opinion mining using sentiment analysis. Decision Analytics, 2(8), Article number: 8. https://doi.org/10.1186/s40165-015-0016-4

History of Spices. (n.d.). McCormick Science Institute. Retrieved April 30, 2023, from https://www.mccormickscienceinstitute.com/resources/history-of-spices

Huotilainen, A., Pirttilä-Backman, A.-M., & Tuorila, H. (2006). How innovativeness relates to social representation of new foods and to the willingness to try and use such foods. Food Quality and Preference, 17(5), 353-361. https://doi.org/10.1016/j.foodqual.2005.04.005

Jabs, J., Devine, C. M., & Sobal, J. (1998). Model of the process of adopting vegetarian diets: Health vegetarians and ethical vegetarians. Journal of Nutrition Education, 30(4), 196-202. https://doi.org/10.1016/S0022-3182(98)70319-X.

Mishra, P. (2022). Zero-shot Text Classification with Hugging Face 🤗 on Gradient. Paperspace Blog. https://blog.paperspace.com/zero-shot-text-classification-with-hugging-face-on-gradient/

MonkeyLearn. (n.d.). Sentiment Analysis. Retrieved May 28, 2023, from https://monkeylearn.com/sentiment-analysis/

Moscovici, S. (1984). Social representations: Studies in social psychology. Cambridge University Press.

Mullin, T. (2020, July 10). DBSCAN Parameter Estimation Using Python. Medium. https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

Onwezen, M. C., & Bartels, J. (2013). Development and cross-cultural validation of a shortened social representations scale of new foods. Food Quality and Preference, 28(1), 226-234. https://doi.org/10.1016/j.foodqual.2012.07.010.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Now Publishers. https://doi.org/10.1561/1500000011

Pattnaik, A. (2020, February 2). Geospatial Clustering: Kinds and Uses. Towards Data Science. Retrieved from https://towardsdatascience.com/geospatial-clustering-kinds-and-uses-9aef7601f386

Pindado, E., & Barrena, R. (2020). Using Twitter to explore consumers' sentiments and their social representations towards new food trends. British Food Journal, 122(12), 3966-3979.

Priy, S. (n.d.). Clustering in Machine Learning. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/clustering-in-machine-learning/

Qin, A. (2016, June 21). Dog Meat and Lychees: A Pairing Meant to Make You Feel Warm Inside. The New York Times. https://www.nytimes.com/2016/06/22/world/what-in-the-world/china-dog-meat-yulin.html

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. https://arxiv.org/abs/2004.09813

Statie, G. (2022, February 16). The Hitchhiker’s Guide to GPT3. Heits Digital. https://heits.digital/articles/gpt3-overview

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. GeoJournal, 78(2), 319-338. https://doi.org/10.1007/s10708-011-9438-2

Stojanovski, D., Strezoski, G., Madjarov, G., Dimitrovski, I., & Chorbev, I. (2018). Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages. Multimedia Tools and Applications, 77, 32213-32242. https://doi.org/10.1007/s11042-018-6168-1

Verdonck, M. (2023). Business Process Management [Lecture and PowerPoint slides]. Ghent University.

Whan-woo, Y. (2019, September 16). Dog meat shunned in South Korea, remains popular in North. The Korea Times. https://www.koreatimes.co.kr/www/nation/2019/09/103_275599.html

Wikipedia contributors. (2023, July 31). Haversine formula. In Wikipedia, The Free Encyclopedia. Retrieved 16:05, August 12, 2023, https://en.wikipedia.org/w/index.php?title=Haversine_formula&oldid=1168079505

Zvornicanin, E. (2023, May 24). What Are Embedding Layers in Neural Networks? Baeldung. https://www.baeldung.com/cs/neural-nets-embedding-layers