GIMA
Geographical Information Management and Applications

# The Online Presence of Liveability

***Using sentiment derived from Twitter messages as an indicator of liveability***

Thesis report
26-05-2021

Mattijs de Haas
5722195
m.j.m.dehaas@students.uu.nl

Supervisor: M. Helbich
Responsible Professor: S. C. M. Geertman

# Abstract

High liveability is associated with benefits like socio-economic equity, inclusive social systems, and improved mental and physical health. Assessing and afterwards effectively improving liveability has been a goal of local, national, and international governments and initiatives. The ambiguity of the concept of liveability complicates its assessment. This study examined the use of residents' sentiment as an indicator of liveability in the Netherlands and explored ways to collect residents' sentiment from location-based social networks.

In this study, eight methods for deriving sentiment from Twitter messages were compared. The best performing method, based on Youden's *J* and percentage of tweets classified, was used to classify the of index dataset of Twitter messages from 2019 located within the Netherlands to sentiment. A total of 1 375 203 tweets was classified, and the resulting sentiment was grouped to the 380 municipalities of the Netherlands. The sentiment scores were compared to the expected liveability patterns, and to an existing liveability index from the Netherlands: the Leefbaarometer 2.0, version of 2018.

A Naïve Bayes' classifier performed the best in the performance assessment, with a Youden's *J* of 0,46. The mean sentiment per municipality showed no similarities to the expected liveability patterns on a national scale, but the expected large regional differences in the peripheral regions were present in the sentiment scores. The mean sentiment and the existing liveability index showed a weak positive relation in the cursory visual analysis, but no statistical relation was found. In conclusion, the sentiment derived from twitter messages in this study does not significantly represent liveability.

Exploring additional sentiment classification methods and using more training data could improve the quality of the sentiment analysis, further solidifying the conclusive strength of research on sentiment and liveability. Another valuable path to follow in future research includes dissecting liveability and analysing the relation of sentiment to its different aspects.

# Contents

# 1. Introduction

## 1.1 Context and problem

Liveable environments are important for community wellbeing, human development, socio-economic equity, and the emergence of inclusive social systems (Caron et al., 2019; Jacobs, 1961; Kashef, 2016; Kovács-Győri, 2019; McGreevy et al., 2020; Tilaki et al., 2014; Wyatt, 2009). These benefits are one of the reasons improving liveability has been the subject of research for over fifty years, and on the agenda of various international initiatives, like the 'Healthy Cities' project started by the World Health Organisation in 1988 (Leidelmeijer & Van Kamp, 2003; McCrea et al., 2020). National and local governments have also tried to improve liveability via policies and projects, to varying degrees of success (Beunen et al., 2020; Kovács-Győri, 2019; McGreevy et al., 2020; Ročak et al., 2016; Ubels et al., 2019).

In order to improve liveability, it is essential to understand what the concept encompasses. The definition of liveability is ambiguous. In urban design, liveability is linked to a mix of residential and commercial land-use, and well-connected but pedestrian-friendly infrastructure. Planning literature generally associates liveability with sustainability, biodiversity, and ecosystems, taking a more holistic approach compared to the approach taken in urban design literature (Kashef, 2016; Ubels et al., 2019). An alternative definition is the quality of the environment in the light of the perception and expectation of the dwellers. This type of definition is commonly used in human-focussed studies (Kovács-Győri, 2019). The concept of liveability will be discussed more in depth in chapter 2.

The inclusion of subjective data is the main difference in assessing liveability between on the one hand design and planning literature, and on the other hand the more people-focussed liveability literature (Kashef, 2016; Kovács-Győri, 2019; Sabbadini & Maggino, 2018). Subjective data are instances of personal interpretations, such as motivations, attitudes, and opinions (Sabbadini & Maggino, 2018). The introduction of these opinions and motivations to a dataset aims at making the results represent the perspective of the population on the contributing factors, instead of describing the state of the environment or standard of living. By assessing the perspective of the population on liveability, the environment can be tailored to their needs (Fu et al., 2019; Kashef, 2016; Kovács-Győri, 2019; McGreevy et al., 2020; Sabbadini & Maggino, 2018; Ubels et al., 2019).

The emotions that residents experience in their environment have been shown to be an indicator of liveability (Ho et al., 2020; Park, 2020; Shafer et al., 2000). However, the collection of data on these emotions can be time-consuming and costly with traditional methods, such as conducting interviews or handing out questionnaires (Alfarrarjeh et al., 2017; Flaes et al., 2016; Javadi & Taleai, 2020; Mitchell et al., 2013). Recently, the popularity of location-based social networks (LBSNs) has been growing as a mean of communication. Researchers seized this opportunity, and the development of tools to gather and analyse the data on these platforms, also called user-generated geo-content (UGGC), has gained popularity as a result. The collection of opinions and attitudes of citizens can now be done in a less time-consuming and more cost effective way (Alfarrarjeh et al., 2017; Flaes et al., 2016; Javadi & Taleai, 2020; Mitchell et al., 2013). UGGC has been used in liveability studies as a complementary data source to objective measures. Sentiment derived from UGGC shows correlations to health, trends in wealth, and levels of education, all contributing factors of liveability (Flaes et al., 2016; Leidelmeijer et al., 2014; Mitchell et al., 2013; Nenko & Petrova, 2019). Following this, a relation between sentiment derived from UGGC and liveability seems likely. However, the use of sentiment derived from UGGC as a direct indicator of liveability has not yet been studied. It offers a powerful, fast and effective approach for collecting opinions on several topics (Schnitzler et al., 2016; Zivanovic et al., 2020). Limitations of this approach for collecting subjective data are the loss of context, and the loss of human interpretation, among others (Javadi & Taleai, 2020; Madhoushi et al., 2015; Zivanovic et al., 2020). Automated classification and interpretation of text is developing rapidly, but indirect textual nuances and sarcasm are still hard to detect. (Giachanou & Crestani, 2016; Javadi & Taleai, 2020).

The importance and impact of contributing factors to liveability are different in rural and urban contexts (Danielaini et al., 2019). The population of the Netherlands is predicted to keep growing, and this growth in an already densely populated country leads to diminishing liveability, mostly in cities (Statistics Netherlands, 2019b). Increase in traffic, decrease in health, and decrease in income equality are three important factors in the diminishing liveability in urban areas (Badland & Pearce, 2019; Khomenko et al., 2020; Municipality of Utrecht, 2016). Large cities that are deemed liveable by indices from popular media do struggle with environmental injustice and health inequality at the same as a result of the crowding (Conger, 2015; Khomenko et al., 2020).

Parallel to this trend of decreasing liveability in highly populated areas, peripheral regions in the Netherlands see a decline in population. This leads to a decline in liveability for different reasons (Beunen et al., 2020; Statistics Netherlands, 2019a, Ročak et al., 2016; Ubels et al., 2019). The decline in population came with loss of jobs and a decline in the concentration of amenities. Local governments have tried to halt the departure of residents in these areas for over fifteen years. However, most strategies were unrealistic and led to a loss of resources, because of ineffective measures (Beunen et al., 2020; Ročak et al., 2016). In more recent strategies, more citizen actors are involved in the process. Incorporating citizen opinions on in the process led to more effective, smaller-scale interventions (Beunen et al., 2020).

### 1.2 Research objective

The main goal of this study is to evaluate the relation between sentiment and liveability in the Netherlands. A secondary, more methodological objective is to explore different approaches of extracting sentiment from UGGC in the Netherlands. To address the problem statement above and these objectives, the  main research question is as follows:

*“To what extent does mean sentiment in municipalities, derived from user-generated geo-content, represent liveability in the Netherlands?”*

Three sub questions have been composed to help operationalise and answer the main research question:

1. What approach of deriving the residents’ sentiment from UGGC gives the most accurate results for content from the Netherlands?
2. To what extent does the spatial distribution of residents' sentiment show similarities to the expected discrepancy of liveability in rural and urban contexts in the Netherlands?
3. To what extent does sentiment of residents show similarities to a liveability index in the Netherlands?

The paper is structured as follows: chapter two discusses the introduced concepts of liveability and sentiment analysis more in depth in a literature review. The third chapter introduces the research area and the data used, and explains the methods used in this study. Chapter four presents the results of the different parts of the analysis, and provides answers to the sub questions. The fifth chapter answers the main question stated above, and provides context for the results.

# 2. Literature review

## 2.1 Defining liveability

In scientific literature, one of the problems of liveability is the ambiguity of the concept (Conger, 2015; Kashef, 2016; Kovács-Győri, 2019). Related concepts like quality of life and standard of living are often used interchangeably with liveability, which can be seen as either one of the causes or one of the results of this ambiguity (Conger, 2015; Leidelmeijer & van Kamp, 2003). The meaning of liveability differs between scientific fields as shown in the introduction, and between rankings and indices published in popular media, like the EIU Liveability Ranking or the Mercer Quality of Living Index (Conger, 2015; Kashef, 2016, Paul & Sen, 2020). In fact, policy makers and researchers use the concept as self-explanatory, and use it as a reference to the living standards and the overall well-being of cities (Paul & Sen, 2020).

The first step towards a clear definition of liveability is done by creating an overview of common definitions used in literature, and distilling the core components. Next, the two mentioned closely related concepts are briefly discussed and separated from liveability. Afterwards, possible contributing factors to liveability are discussed, and how the importance of these factors changes in highly urbanised areas and depopulating areas. The last section discusses how to operationalise these contributing factors.

In table 2.1, a set of definitions of liveability is presented. These definitions will be compared and discussed. Afterwards, the definition used in this report is presented, distilled from several of the definitions from table 2.1. This collection of definitions is by no means comprehensive, but covers the interpretations of liveability in literature.

The relation between people and their environment is a core component of all definitions from table 2.1. The definitions proposed in Conger (2015), Shafer et al. (2000), and Zivanovic et al. (2020) are broad and give little direction apart from this relation. The way liveability is defined in Antognelli & Vizarri (2017) separates a subjective part of liveability, the characteristics of persons, and an objective part of liveability, the qualities of landscapes. It does not indicate a directional relation between the subjective and objective parts however. The more specific definitions represent two different directions of this human-environment relation: the environment must be of high quality to make life as pleasant as possible (e.g. Badland et al, 2014; Khomenko et al., 2020; Treija et al., 2020), and the environment must meet the expectations of the residents (e.g. Kovács-Győri, 2019; Leidelmeijer & van Kamp, 2003; Veenhoven, 1996). Liveability using the first perspective assumes liveable environment is objective and the same for everyone, and evaluates the happiness of residents based on the environmental quality. This definition of liveability is very close to what quality of life is defined as in the next paragraph. The way residents interact with and give meaning their environment is based on personal characteristics, making an assessment based on the residents is more meaningful than an assessment based on the environment for improvements of liveability (Badland & Pearce, 2019; Leidelmeijer & van Kamp, 2003). In section 2.2 this is further explained. The second perspective assumes a liveable environment is different for different people and communities, and is more subjective. The following definition is used in this study: liveability is to what extent the environment meets the needs and expectations of its residents. It is based on the definitions of Antognelli & Vizzari (2017), Leidelmeijer & Van Kamp (2003), Kovács-Győri (2019), and Veenhoven (1996). This definition is chosen because it gives a clear direction for assessing liveability. It includes the subjective perspective of the residents and objective characteristics of the environment. The directional nature of the relation is indicated by the requirement of the environment to provide for its residents.

| Article | Definition |
|---|---|
| Veenhoven (1996, p. 7) | “Liveability of a nation can be defined as the degree to which its provisions and requirements fit with the needs and capacities of its citizens.” |
| Shafer et al.<br>(2000, p. 178) | “[…] liveability, in this case the interaction between a community and the environment.” |
| Leidelmeijer & van Kamp (2003, p. 59) | “[…] whether the living environment meets the wishes and needs that are set by the residents and if so, to what extent these wishes and needs are met.” |
| Badland et al. (2014, p. 65) | “Liveable cities are socially inclusive, affordable, healthy, safe and resilient to the impacts of climate change. They have attractive built and natural environments. Liveable cities provide a choice and opportunity for people to live their lives, and raise their families to their fullest potential.” |
| Conger<br>(2015, p. 6) | “[…] the most liveable city is not necessarily the ‘best’, [it is] simply the least challenging in which to live.” |
| Antognelli & Vizzari (2017, p. 704) | “Liveability theory assumes that the perceived quality of life is dependent on both subjective characteristics of persons and objective qualities of landscapes.” |
| Kovács-Győri<br>(2019, p. 290) | “[…] the quality of the person-environment relationship in the urban context, concerning the needs and expectations of the residents towards the urban environment.” |
| Zivanovic et al.<br>(2020, p. 239) | “[…] the relation between people and their living environment.” |
| Khomenko et al.<br>(2020, p. 12) | “[…] the current liveability definition [encompassing stability, access to healthcare, culture and environment, education and infrastructure] should consider more comprehensively health and wellbeing, environmental, and socioeconomic indicators.” |
| Treija et al. (2020, p. 17) | “Generally, liveability is defined as the sum of the factors that add up to a community’s quality of life – including the built and natural environments, economic prosperity, social stability and equity, educational opportunity, and cultural, entertainment and recreation possibilities.” |

Table 2.1: Different definitions of liveability found in literature.

This paragraph discusses two concepts that are related to liveability, and separates their definitions from the definition of liveability presented above. Liveability represents one of the angles used to look at the relation between humans and their environment. Concepts that represent other ways to look at this relation are often used as a synonym of liveability, but have a different meaning. The first of the closely related concepts is standard of living. Most definitions of this concept include these two core parts and their relation: the wellbeing of residents and objective quality of the physical environment (Kashef, 2016; Leidelmeijer & van Kamp, 2003). Indices and studies that use objective measures and demographics to evaluate liveability actually evaluate standard of living (Conger, 2015; Kashef, 2016; Paul & Sen, 2020). The difference between both, according to the definitions of liveability presented above, is that in liveability the perspective of people is included in the evaluation of the environment.

Quality of Life (QoL) is the second closely related concept to be discussed. Both QoL and liveability are concerned with the relation between humans and their environment, the environment being social, built, natural, economic and cultural (Conger, 2015; Leidelmeijer & van Kamp, 2003; Zivanovic et al., 2020). However, QoL measures the effect of the environment on the lives of its dwellers, and liveability measures to what extent the environment meets the expectations of its dwellers (Conger, 2015; Kovács-Győri, 2019; Leidelmeijer & van Kamp, 2003). Following this distinction, standard of living is a direct contributor to QoL, while in liveability standard of living is included through the eyes of its dwellers.

## 2.2 Contributing factors of liveability

In the previous paragraphs, liveability is introduced as an ambiguous concept. This ambiguity is one of the reasons why liveability is hard to assess, and why local and national governments in Europe and Australia struggle to manage it (Beunen et al., 2020; Kovács-Győri, 2019; McGreevy et al., 2020; Ročak et al., 2016). Another reason why liveability is hard to assess correctly, is the complexity of contributing factors (Badland & Pearce, 2019). This section discusses different contributing factors in general, as well as the importance of various factors in highly urbanised areas and areas dealing with depopulation.

The way liveability is defined in this study, the contributing factors can be categorised in two groups: the state of the environment, and factors that influence the needs and expectations of the environment of the residents. The state of the environment can be assessed through objective measures, like employment opportunities, housing price and quality, access to and quality of education, proximity and quality of amenities, and proximity of greenery. An advantage of using objective measures is the accessibility and ease of data collection (Badland et al., 2014; Conger, 2015; Kovács-Győri & Reinel, 2017; Leidelmeijer & van Kamp, 2003; Shafer et al., 2000).

The perspectives and expectations of residents are formed by personal character traits and values, financial possibilities, gender, education, health, etc. (Badland & Pearce, 2019; Badland et al., 2014; Leidelmeijer & van Kamp, 2003; Namazi-Rad et al., 2012). These factors influence how an individual interacts with the surroundings and what an individual needs and expects, and can be assessed using objective or subjective data (Badland & Pearce, 2019; Badland et al., 2014; Leidelmeijer & van Kamp, 2003). One of the weaknesses of using objective data to assess liveability is the weighting of the different measures. The weighting should reflect the preferences of the residents, and cannot be created without subjective data (Kovács-Győri & Reinel, 2017; Leidelmeijer & van Kamp, 2003). Subjective data on the other hand needs objective data to provide context to reflect on the difference between stated preference and revealed preference (Leidelmeijer et al., 2014). The relation between liveability, the environment, and needs and expectations of residents is shown in figure 2.1.

The contributing factors of liveability are the same in highly urbanised areas and areas with declining population, but the relative importance of the factors to the residents changes. Urban residents tend to put more emphasis on recreation, transport and quality of amenities (Danielaini et al., 2019; Lapointe et al., 2019). Residents of more peripheral regions and regions with shrinking populations are inclined to assign more weight to quality and proximity of greenery, and employment opportunities (Danielaini et al., 2019; Gieling & Haartsen, 2016). The needs and expectations of residents differs between these areas, leading to a different assessment of the environment (Danielaini et al., 2019).

Highly urbanised areas often have a higher concentration in amenities, closer and better educational institutes, but also have to deal with more traffic, more noise pollution, less green space, and the increase in concentration of public and private services sometimes does not suffice to meet the needs of current and additional residents (Howley et al., 2009; McCrea & Walters, 2012). Residents in highly urbanised areas often experience a lack of community involvement, and as a result reduce their social communication, and walking and cycling time in public spaces (Howley et al., 2009; Tilaki et al., 2014). An increase in population density in these areas is often the result of urban consolidation. Urban consolidation is a planning strategy aimed at redesigning the space within the city boundaries to facilitate more residents, as opposed to expanding the city boundaries to do so. However, this leads to the dilution or loss of local identities and lifestyles, further decreasing community feeling and social interaction (Beattie & Haarhoff, 2018; McCrea & Walters, 2012).

Decline in population is often accompanied by an increase in housing abandonment, social problems, crime, and a decline of quantity and quality of educational institutes (Beunen et al., 2020; Hollander, 2011; Ubels et al., 2019). These are all directly associated with liveability (Badland & Pearce, 2019; Badland et al., 2014; Conger, 2015; Hollander, 2011; Kovács-Győri & Reinel, 2017; Leidelmeijer & van Kamp, 2003; Shafer et al., 2000). There are different ways the decline in liveability can be mitigated or turned around. Public redevelopment, focussed on improving economic conditions through development plans that emphasize controlled shrinkage in smaller but liveable places, and improving the physical realm of an area through re-greening once built-up areas and depopulating run-down neighbourhoods are the most prevalent approaches governments take (Blakely & Bradshaw, 2002, as cited in Hollander, 2011; Weichmann & Bontje, 2015). In the US, Hollander (2011) showed a large difference in the trend of perceived liveability between different shrinking cities, and Weichmann & Bontje (2015) state that restructuring depopulating urban regions is one of the most challenging tasks for Europe's cities in the upcoming years. The strategies that resulted in mitigation of the decline of liveability in shrinking cities focussed on improving the city for its current residents, instead of attempting to grow the city (Beunen et al., 2020; Hollander, 2011; Ročak et al., 2016).

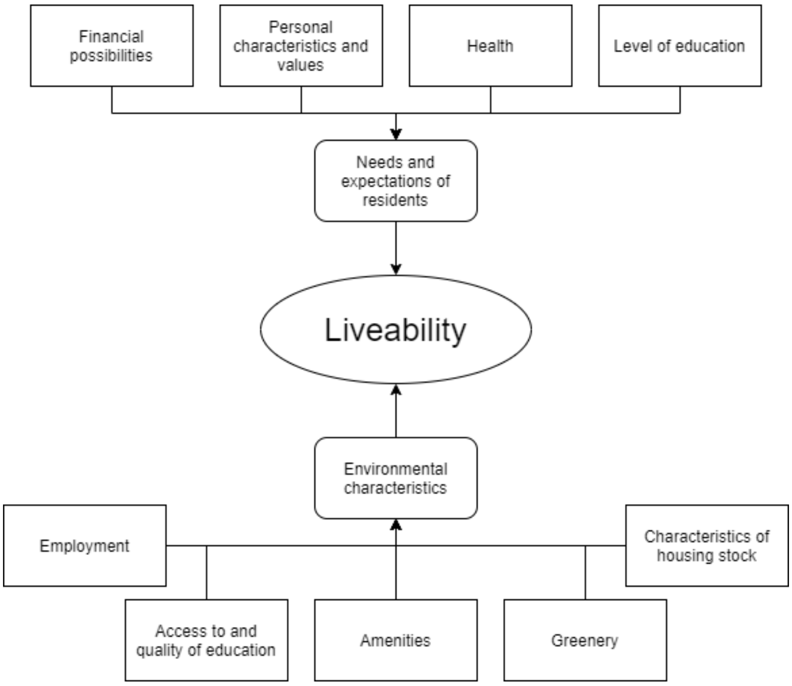

Financial possibilities
Personal characteristics and values
Health
Level of education
Needs and expectations of residents
Liveability
Environmental characteristics
Employment
Characteristics of housing stock
Access to and quality of education
Amenities
Greenery

*Figure 2.1: Relation between liveability, environmental and individual characteristics.*

### 2.3 Operationalisation

Indices like the Global Liveability Index by the Economist Intelligence Unit (EIU) and the Mercer Quality of Living Index use similar objective indicators for measuring the quality of the environment, but they come to different conclusions on what the most liveable environment is. This is because of the difference in the weights assigned to the indicators (Badland et al., 2014; Conger 2015; Khomenko et al., 2020; Paul & Sen, 2020). In order to capture the liveability of different places with static objective data, the weights of the indicators have to be adjusted to the preferences of the local population (Beunen et al., 2020; Clarke & Cheshire, 2018; McCrea & Walters, 2012; Ročak et al., 2016). The weighting of the indicators can be based on objective and subjective data. The indices from the EIU and Mercer survey their employees and base the weights of the different categories on their opinions. Monocle's ranking is based on a more narrow base: the opinions of the editorial board (Conger, 2015). An inherent issue of this approach is the bias of the weights towards the opinion of a non-representative sample of the population. All three of the indices assess cities all over the world, while their weights are based on a mostly Western, elitist perspective (Conger, 2015; Kotkin, 2009). The Leefbaarometer 2.0 constructed by Leidelmeijer et al. (2014) is a liveability index for the Netherlands. It uses both subjective and objective data as dependent variables in a regression analysis to weight their indicators. A national survey is the basis for their subjective dependent variable, representing the stated preferences of the residents. The objective dependent variable is based on hedonic residential property price, representing the revealed preference of the residents. The two separate models based on the subjective and the objective data are used to produce the final liveability index. This index takes the opinions of a representative sample of the population into account, as well as an objective measure to validate the subjective weights.

Subjective data are also used as a direct indicator of liveability, not only as a way to weight objective indicators (e.g. Flaes et al., 2016; Javadi & Taleai, 2020). Conflicts between expectations of the environment and the factual environment lead to emotions, positive or negative (McGreevy et al., 2020). These emotions can be indicators of liveability, and the adjustment to personal preference is included in the emotions themselves (Ho et al., 2020; Park, 2020; Shafer et al., 2000). Not all emotions are directly related to the environment or residents' satisfaction with it, but satisfied residents do feel and express positive emotions more often (Caron et al., 2018; Mitchell et al., 2013; Park, 2020). One of the major disadvantages of subjective data relative to objective data are the time and resources needed for the collection and analysis. It is more time and resource consuming to match the temporal and spatial scale of objective data with subjective data (Alfarrarjeh et al., 2017; Flaes et al., 2016; Javadi & Taleai, 2020; Mitchell et al., 2013; Sabbadini & Maggino, 2018). Collecting and analysing UGGC as the subjective data can compensate for this gap, but also introduces its own challenges.

The emotions conveyed in the content from the LBSNs can be extracted using sentiment analysis. Sentiment analysis is a discipline within the natural language processing. Natural language processing is the scientific field concerned with interactions between computers and human language (Devika et al., 2016; Fathy et al., 2017). The aim of sentiment analysis is to extract opinionated or emotional information from shared text (Devika et al., 2016; Dey et al., 2016; Ghiassi et al., 2013; Giachanou & Crestani, 2016; Manad et al., 2019). It has been applied to texts at various scales: on a document as a whole, on sentences, and also on phrases. (Agarwal et al., 2011; Aisopos et al., 2011; De Smedt & Daelemans, 2012a; Liu & Zhang, 2012; Tripathy et al., 2015). Application at the latter two scales is often an intermediate step towards sentiment analysis of a document as a whole (Liu & Zhang, 2012).

Twitter is a popular LBSN, with a rapidly growing user base (Flaes et al., 2016; Ghiassi et al., 2013; Javadi & Taleai, 2020; Mitchell et al., 2013). Not only is it popular among users, it has also seen a rise of popularity among researchers as a source of data for sentiment analyses, for both academic and commercial purposes (Dave et al., 2021; Giachanou & Crestani, 2016; Javai & Taleai, 2020; Jianqiang & Xiaolin, 2017). It has been used for the evaluation of quality of life and liveability in academic research, and in comparison to other LBSNs performed the best. This is likely because of the high frequency and limited size of posts, forcing users to write concise and clear messages (Antonakaki et al., 2021; Flaes et al., 2016; Javadi & Taleai, 2020; Mitchell et al., 2013). Another advantage of Twitter is the relative ease of data collection and processing compared to other popular LBSNs, due to the easily accessible Twitter API and its simple data structure (Curini et al., 2015). The sentiment analysis of tweets is not without its challenges however, with most of them relating to interpretation of the text in its context (Giachanou & Crestani, 2016; Javadi & Taleai, 2020; Liu & Zhang, 2012). Pre-processing the twitter data is a step that significantly improves the performance of the sentiment analyser. It is done by reducing the flaws in the use of language often present in online text resources (Go et al., 2009a; Jianqiang & Xiaolin, 2017; Liu & Zhang, 2012). Section 3.2 elaborates on the implementations in this study.

## 2.4 Conceptual model

Figure 2.2 is a schematic representation of the theoretical relations between the emotions of residents, liveability, and the contributing factors. Based on the theoretical relation between liveability and some of its aspects, and the emotions of residents presented earlier this chapter, this study uses emotions from residents derived from Twitter content as an indicator for liveability. The differences between highly urbanised areas and depopulating areas are present in the environmental characteristics. These differences can both positively and negatively impact liveability and emotions of residents, depending on the needs and expectations. The dotted lines indicates the relation between liveability and the contributing factors are part of the theoretical framework, but are not taken into account in the operationalization.

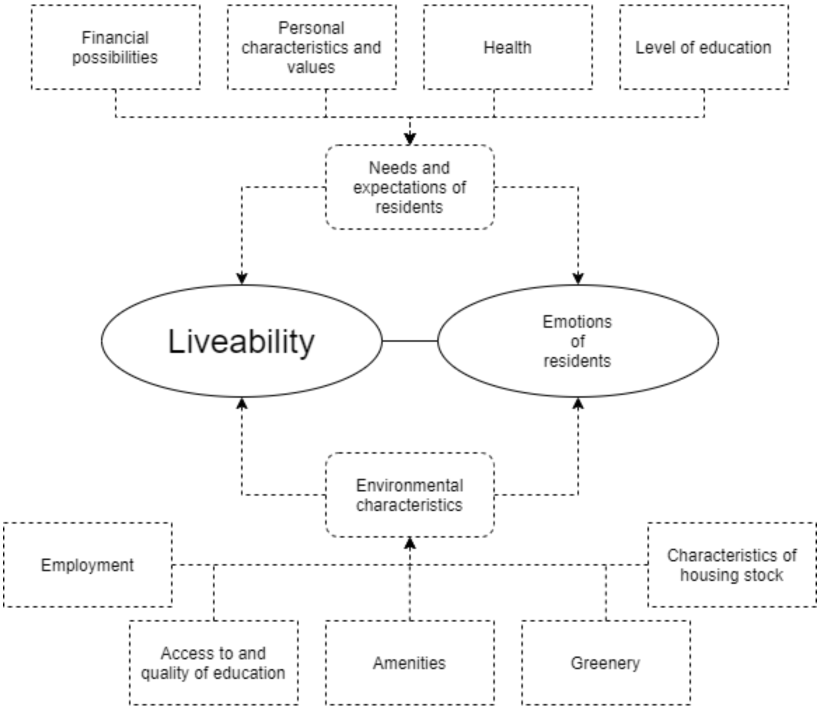

Financial possibilities
Personal characteristics and values
Health
Level of education
Needs and expectations of residents
Liveability
Emotions of residents
Environmental characteristics
Employment
Characteristics of housing stock
Access to and quality of education
Amenities
Greenery

*Figure 2.2: Conceptual model.*

# 3. Methods

## 3.1 Research area

The research area of this study is the Netherlands. The Netherlands can be divided into 380 municipalities. The western regions of North-Holland, South-Holland and Utrecht are more densely populated and have a higher concentration of amenities compared to the other regions, and have the highest percentages of built-up area. In the East and North, the population density is lower, amenities are more spread out, and forests and agricultural area take up more of the total area (Statistics Netherlands, 2021 March 31). Figure 3.1 shows the research area.

The gaps in liveability have widened in the Netherlands between 2016 and 2018, within cities, regions, and the nation as a whole (Leidelmeijer et al., 2019). The value of using sentiment in liveability assessments has recently been researched in the Netherlands at city level by Flaes et al. (2016), Amsterdam to be specific. This assessment has not been done at national level yet. Furthermore, the six areas that require attention regarding liveability according to Leidelmeijer et al. (2019) are located both in the western, more urbanised areas as well as the more peripheral regions.

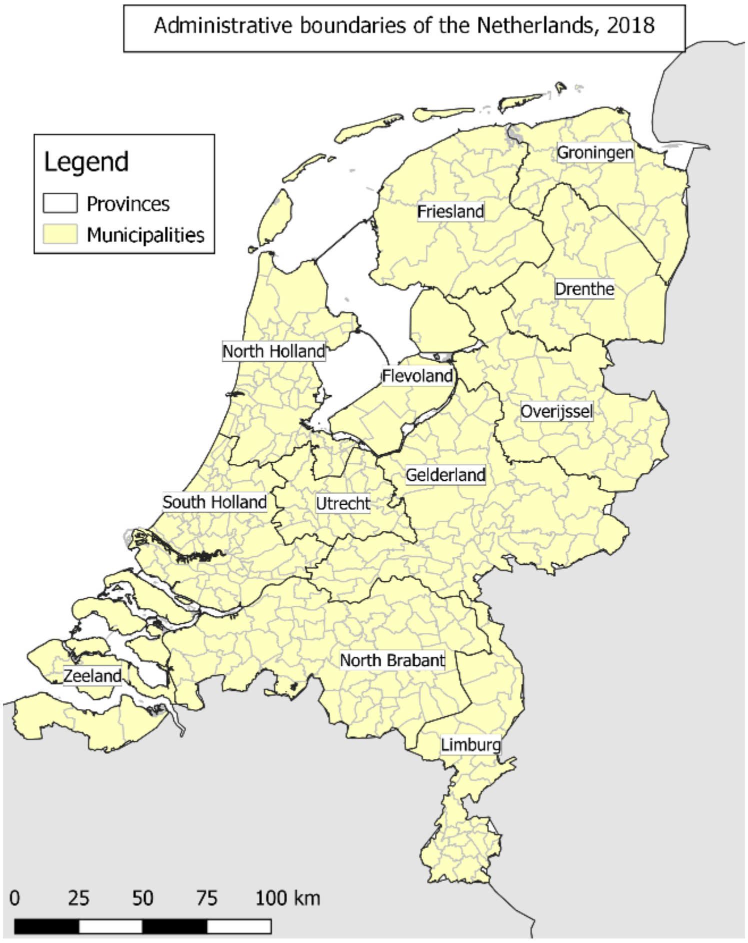


*Figure 3.1: Provincial and municipal boundaries of the Netherlands, 2018.*

### 3.2 Data

This section covers the collection of the data, explores the characteristics of the data, and introduces the pre-processing steps that are taken to prepare the data for the sentiment analysis. The main source of data for this research is UGGC, specifically from Twitter. The secondary dataset for this research is an existing liveability index: the Leefbaarometer 2.0, version of 2018.

The Twitter data that are used in this study are from 2019, and are gathered using SNScrape and Hydrator, a python package for accessing historical tweets and a desktop application for gathering tweet data based on tweet IDs respectively. All available and eligible *tweets*, status-messages on Twitter, sent within the research area are used in this research, as all tweets can have an emotional charge or contain an opinion (Dave et al., 2021; Sarma et al., 2019).

Tweets from outside the research area are filtered out. Tweets from '*bots'*, automated users, are also filtered out as much as possible, as they contain no emotions or sentiment related to a living environment (Alfarrarjeh et al., 2017; Varol et al., 2017). Bots can be detected by analysing their network or their behaviour (Ferrara et al., 2016). For this study, bots are excluded based on their behaviour: users that tweet more than 15 times per day are labelled as bots and the tweets from these users are deleted (Alfarrarjeh et al., 2017; To et al., 2017). This will not detect all bots, but will significantly decrease the impact (Shao et al., 2018).

Furthermore, the text needs to be translated to a single language before it can be used as input for a classifier. GoogleTrans, a python package for accessing the Google Translate interface, is used to recognise the language of the tweet and translate the tweets to a single language. The most prevalent language in the tweets is Dutch with 42% of the tweets, followed by English with 25%. The performance of the sentiment classifiers is tested on both these languages. Dutch is chosen because the amount of tweets that need to be translated is lowest, minimising the impact of this step on the performance of the classifiers (Agarwal et al., 2011; Basiri & Kabiri, 2017; Manad et al., 2019). English is chosen as a second language, because most tools for NLP are compatible with this language, while still keeping the impact of translation on the performance limited (Agarwal et al., 2011).

The main dataset of tweets from 2019 in the Netherlands consists of 3,75 million geo-tagged tweets. 59% of the gathered tweets from the dataset are located within the Netherlands, and a further 39% of the tweets from the Netherlands were filtered out in the exclusion of automatically generated tweets. This leaves 1 375 203 usable tweets, 37% of the starting dataset.

The location data of tweets can be very precise, if the user pairs the current coordinates of their device to the tweet. Just over 9 percent of the tweets from the dataset contains the coordinates of the location user the moment the tweet was sent. Other types of tweet location are place coordinates and user location. Place coordinates and user location are less accurate however than tweet coordinates, as they refer to named locations chosen by the user. This reference can be to a restaurant, but also to a city or country (Dredze et al., 2013; Zivanovic et al., 2020). The remaining 91 percent of tweets in the dataset are geo-tagged with place coordinates. Most of the place coordinates in the dataset refer to a city. This means the municipal scale is the largest meaningful geographical scale of analysis.

The main dataset is split into an index dataset, a training dataset, and a test dataset. The index dataset is the dataset that is used to determine the sentiment per municipality. The training dataset is used to train the learning algorithms, and consists of 570 tweets. In previous performance assessments, this was enough to train a learning-based classifier to an accuracy of 70-75% for a classification into two classes (Abdelwahab et al., 2015; Needham et al., 2006). The actual performance of the classifiers is shown in the performance assessment in section 4.1. The test dataset consists of 235 manually classified tweets after pre-processing, and is used to test the performance of the different approaches presented below. The split of training data and test data is 70/30. The approaches and performance measures are explained in section 3.3. The approach that performs best on the test dataset is used as the classifier for the index dataset.

To evaluate the results of the subjective liveability assessment, the sentiment index constructed in this study is compared to an objective liveability index, the Leefbaarometer 2.0, version of 2018. This index covers the whole research area, and is available at the municipal scale. It is based on open geographical and statistical data from the Netherlands. The five main domains of this index are: characteristics of housing stock, physical environment, amenities, residents, and safety. 400 indicators from the main domains were tested for qualitative integrity and quantitative relevance. The test for qualitative integrity is done on the basis of five criteria. The indicators have to: be explainable; have data available with nationwide coverage; have data available at a low geographical scale; have data available in a continuous time series; and be reliable. If an indicator passed all five criteria, it was taken into account in the test for quantitative relevance.

The quantitative test resulted in two models: an ‘stated preference’ model, where the result of a national liveability survey is the dependent variable, and a ‘revealed preference’ model, where the hedonic residential property price is the dependent variable. These models are used to assign weights to the 100 indicators, based on their predictive power. The results of the models are standardized to Z-scores. The mean of the standardized scores is classified to ordinal classes ranging from poor to excellent. This classification is used as the definitive Leefbaarometer score. A more in depth coverage of the process and the exact calculations used can be found in Leidelmeijer et al. (2014).

The number of municipalities per class in the index of 2018 are shown in table 3.1. Because the more extreme classes contain little or no observations, the classes are reclassified for this research. Classes lower than the median class ‘good’ are grouped and classes higher than the median class are grouped. Table 3.2 shows the new classes and number of municipalities in these new classes.

| Leefbaarometer class | Number of municipalities |
|---|---:|
| Poor | 0 |
| Strongly insufficient | 0 |
| Insufficient | 0 |
| Weak | 3 |
| Acceptable | 2 |
| Satisfactory | 103 |
| Good | 170 |
| Very good | 90 |
| Excellent | 12 |

*Table 3.1: Original classes of the Leefbaarometer 2.0 index with the corresponding number of municipalities from the 2018 version*

| Class | Number of municipalities |
|---|---:|
| Low | 108 |
| Medium | 170 |
| High | 102 |

*Table 3.2: New classes with the corresponding number of municipalities from the Leefbaarometer 2.0 index, 2018 version.*

### 3.3 Analytical approach

The text from the collected tweets is subjected to a sentiment analysis. In its core, sentiment analysis methods compare characteristics of the text to some form of prior knowledge on what these characteristics mean. The characteristics of the text can be words, patterns and networks of words, or patterns in groups of letters for instance. The prior knowledge can be a preconstructed lexicon, or the training data of a supervised learning algorithm. Techniques for sentiment analysis can be grouped in three categories: lexicon-based approaches, learning-based approaches, and hybrid approaches (Fathy et al., 2017; Javadi & Taleai, 2020). The first involves application of a sentiment lexicon of emotion-related positive or negative terms to evaluate the text. The lexicon consists of a list of words, each paired to a sentiment or emotion score. The words of the text are matched to the scores in the dictionary. The sentiment score of the text is a sum of the sentiment score of each word in the tweet (Fathy et al., 2017; Jianqiang & Xiaolin, 2017).

Learning-based methods can be divided into supervised and unsupervised learning methods. Supervised learning algorithms are trained by feeding it a pre-labelled dataset. The algorithm uses the examples and patterns from the pre-labelled dataset to analyse the new dataset. Unsupervised learning algorithms are not trained with a pre-labelled dataset, but instead finds patterns in the new dataset, and is not bound to predetermined categories (Fathy et al., 2017; Madhoushi et al., 2015). Supervised learning algorithms are used more commonly in NLP and twitter analysis specifically. Unsupervised learning algorithms require relatively more training data. Furthermore, they often produce incoherent topic categories, because the objective functions of topic models do not match with human judgments (Madhoushi et al., 2015). Even though unsupervised learning approaches are more intricate and may become better as they are further developed, supervised learning methods still outperform unsupervised learning methods in sentiment analysis (Kapočiūtė-Dzikienė et al., 2019; Rogers et al., 2018).

Hybrid methods are less easily defined, as this can be any combination of learning-based and lexicon-based approaches. Two possibilities are the instances from the lexicon are used as (a part of) the training dataset for supervised learning algorithms, or learning-based methods are used to expand an existing sentiment lexicon if words are not found in said lexicon (Fathy et al., 2017; Javadi & Taleai, 2020).

For each of the two basic sentiment analysis approaches, lexicon-based and learning-based, two methods are tested. For the hybrid methods, all four possible combinations of the lexicon-based and learning-based methods are tested. Below, the main characteristics of all eight methods are briefly discussed.

*Lexicon-based methods*

Two lexicon-based methods are evaluated: EmoWordNet, an English lexicon constructed by Badaro et al. (2018), and a Dutch lexicon constructed by De Smedt & Daelemans (2012a). The Dutch lexicon is chosen so most tweets do not have to be translated, as most tweets are already in Dutch, minimising the impact of this pre-processing step (Agarwal et al., 2011). The

English lexicon is chosen because it is extensive and performs well, and English is the second most occurring language in the dataset, also limiting the impact of errors during translation (Badaro et al., 2018; Agarwal et al., 2011).

For the implementation of the lexicon-based and learning-based methods, this research uses Pattern. This is a python package for natural language processing and machine learning. It focusses on ease of use and allows for importing custom lexicons and several learning algorithms (De Smedt & Daelemans, 2012b). The first of the two lexicons used is EmoWordNet. This is an English lexicon, designed for detecting and classifying emotional or opinionated text in social media (Badaro et al., 2018). It expands on an existing emotional lexicon published by Staiano and Guerini (2014), called DepecheMood. The terms in DepecheMood and their cognitive synonyms in English Wordnet, a large lexical database, were paired. Based on the scores of their DM pairings, each term in English Wordnet received an emotional score (Badaro et al., 2018; Fellbaum, 2006). The lexicon consists of around 67000 terms, annotated with 8 emotion scores.
The second keyword database used in this study is a Dutch subjectivity lexicon. This lexicon has been constructed by doing a manual annotation of online product reviews and a machine learning based expansion of the manually generated initial lexicon, and only classifies adjectives. The automatic expansion was done by pairing new adjectives from a more extensive database, and picking their nearest neighbours. These nearest neighbours are determined by how many common nouns the adjectives often directly precede. For instance, '*fantastisch'* (fantastic) was close neighbours with '*geweldig'* (great), but also with 'horrible' and 'electoral'. After the process, the neighbours were manually checked to reduce antonymy (horrible), noise (electoral), and word-sense disambiguation. A second expansion was done in a similar way to the construction of EmoWordNet. By using the cognitive synonyms from CORNETTO, a Dutch WordNet derivative, of the unclassified adjectives were given a score based on the score of their classified related adjectives (De Smedt & Daelemans, 2012a). In order to use the Dutch lexicon, the text has to be tokenized, and all but the adjectives have to be filtered out. One of the main advantages of the lexicon-based methods is that the scores are based on  or verified by human interpretation of the words. This means the scores are very likely to be correctly classified in their context (Badaro et al., 2018; De Smedt & Daelemans, 2021a; Fellbaum, 2006). The main weakness is the inability to adapt to terms that do not occur in the dictionary, meaning it will not be able to classify these words. The more extensive the lexicon is, the lower the impact of this weakness on the performance (Baid et al., 2011; Devika et al., 2016; Dey et al., 2016; Fathy et al., 2017; Javadi & Taleai, 2020).

*Learning-based methods*
For this study, Naïve Bayes' (NB) and K-Nearest Neighbours (K-NN) methods are tested as the supervised learning-based approaches. NB methods consistently outperform other supervised learning-based approaches in sentiment analysis. K-NN methods are among the top performers in other comparative studies as well, and works well in hybrid methods (Go et al., 2009b; Jianqiang & Xiaolin, 2017; Kapočiūtė-Dzikienė et al., 2019; Rogers et al., 2018).

The NB method is based on Bayes' law, and estimates the probability of an event, based on prior knowledge or conditions (Dey et al., 2016). In a NB classifier, the estimation is done by comparing an array of independent characteristics of the input to the training data (Baid et al., 2011; Devika et al., 2016; Dey et al., 2016). In sentiment analysis the event is the sentiment of the document falling under a sentiment class (e.g. positive or negative) (Baid et al., 2011; Devika et al., 2016; Dey et al., 2016). It is based on the following equation:

$$P(A|x) = \frac{P(x|A) * P(A)}{P(x)} \qquad (1)$$

Where *P(A|x)* is the probability that the document falls under class *A* given predictor *x* is right, *P(x|A)* is the probability that predictor *x* is right given the document falls under class *A, P(A)* is the probability that the document falls under class *A*,

$$P(x) = P(x|A) * P(A) + P\ (x|B) * P(B) \qquad (2)$$

Where *P(x)* is the probability that predictor *x* is right, *P(x|B)* is the probability that predictor *x* is right given the document does not fall under class *A*, and *P(B)* is the probability that the document does not fall under class *A*. The predictors are formulated by the algorithm analysing all characteristics of the text, and how these characteristics behave in the given classes. A major strength of this method is that it shows a similar performance to or outperforms highly sophisticated classification methods, even more so with a small training dataset (Baid et al., 2011; Dey et al., 2016).

The second learning-based classifier tested in this research is a K-NN classifier. This type of classifier is based on the principle that words that often occur close to each other in a sentence are likely to have similar sentiments. Words are considered close to each other if they are within *k* words of each other. The words in the classified documents receive a value, based on how often they occur in positive and negative documents, building a dictionary. The documents in the dataset that needs to be analysed receive a score based on the scores from the dictionary. If a word does not exist in the dictionary, the scores from *k* nearest neighbours are taken and assigned to the unknown word (Baid et al., 2017; Devika et al., 2016; Dey et al., 2016). One of the advantages of this method is that each word receives a score, whether they occur in the training dataset or not. One of the weaknesses is that the performance of the method is highly dependent on the size of the training dataset (Baid et al., 2017; Devika et al., 2016; Dey et al., 2016; Fathy et al., 2017; Huq et al., 2017; Javadi & Taleai, 2020).

*Hybrid methods*

The hybrid methods used in this study work as follows: the lexicon classifies all tweets it is able to. The classified tweets used to expand the training dataset for the learning-based methods, and the unclassified tweets are fed into the learning-based algorithm to be classified. Hybrid methods are designed to cover the weaknesses of the methods they are composed of, in this case the lexicon-based methods and learning-based methods (Fathy et al., 2017). The main weakness of lexicon-based methods is their inability to adapt. This means the lexicons will likely not be able to classify all tweets (Baid et al., 2011; Devika et al., 2016; Dey et al., 2016; Fathy et al., 2017; Javadi & Taleai, 2020). The main weakness of learning based methods is the dependence of their performance on the size and accuracy of the training dataset (Baid et al., 2017; Devika et al., 2016; Dey et al., 2016; Fathy et al., 2017; Huq et al., 2017; Javadi & Taleai, 2020).

The impact of using a hybrid method with the EWN lexicon is likely limited. This is because this lexicon is extensive and all types of words can be used as input, meaning it can be assumed all tweets can be classified by this lexicon. For the lexicon constructed by De Smedt & Daelemans (2021), the impact of using a hybrid method is larger. All but the adjectives have to be filtered out for this lexicon. Not all tweets contain adjectives, meaning not all tweets will be classified by this lexicon. Adding a learning-based classifier to classify these tweets increases the coverage, as these methods classify tweets based on probability (Dey et al., 2016).

Learning-based methods also benefit from the implementation in hybrid methods. This is because their training data is more extensive than without. This means their performance will increase, assuming the added training data is accurate (Baid et al., 2017; Devika et al., 2016; Dey et al., 2016; Fathy et al., 2017; Huq et al., 2017; Javadi & Taleai, 2020).

*Performance assessment*

For the performance assessment, the tweets from the test dataset and training dataset are manually scored for sentiment. For the classification using the lexicon-based methods, a threshold score is needed to separate the classes. Tweets that are assigned a score above this threshold score is seen as positive and vice versa. This threshold sentiment score is also called the threshold polarity. The threshold polarity is one of the parameters that is dialled in in this performance assessment. The tested values range from -0,3 to 0,3 and is changed in steps of 0,05.

For the learning-based methods, different types of input are tested. The first type is plain text, the same as the input for the lexicons. The second type is *n*-gram. An *n*-gram is a selection of *n* amount of following features from the starting feature. Features can for instance be characters, words, or word groups. Word *n*-grams and character *n*-grams are tested with n = 1 to n = 5. For NB, the baseline and method parameters are the default values, as for the K-NN the baseline, *k*, and distance. Each of the proposed methods and parameters is tested for accuracy, $F_1$-score, Youden's *J*, and wat percentage of the tweets is classified. These are calculated as described in formula 3-8.

$$accuracy = \frac{true\ positive + true\ negative}{total} \tag{3}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{4}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{5}$$

$$specifity\ = \frac{true\ negative}{true\ negative\ + false\ positive} \tag{6}$$

$$F_1 = \frac{precision * recall}{precision + recall} \tag{7}$$

$$J = recall + specificity - 1 \tag{8}$$

The percent classified tweets, accuracy, $F_1$-score, and Youden's *J* are the measures that will be taken into account for this performance assessment. The percentage of tweets that is classified is used to measure how much of the available data is useful for the classifier. A high percentage means the coverage of municipalities will be better, so this is also taken into account. Accuracy shows what part of the total tweets is appointed to the right class by the classifier. An accuracy of 0 means no tweet has been correctly identified, 1 means all tweets are correctly identified. The $F_1$-score reflects the performance in and balance between precision and recall. Precision shows the ratio of correctly identified positive results to all positive results, and recall the ratio of correctly identified positive results to all samples that should have been identified as positive. The score can be between 0 and 1, with scores closer to 1 indicating higher performance. The $F_1$-score does not take the performance of the classifier in identifying negative results however. Youden's *J* captures the performance of the classifier in identifying both positive and negative results by adding recall and specificity. Specifity is the ratio between correctly identified negative results to all samples that should have been identified as negative. Youden's *J* can show values between -1 and 1, and the higher score indicates a higher performance. A high Youden's *J* indicates that the performance on negative and positive tweets is balanced, and a low *J* indicates a bias towards either positive or negative tweets. For this study, the performance in identifying both positive and negative tweets is important. This is why Youden's *J* is the leading measure in selecting the right parameters and selecting the classifier for the index.

After the most suitable method has been chosen based firstly on Youden's *J*, and secondarily percentage classified, accuracy, and the $F_1$-score, the method will be applied to the index dataset. The derived sentiment scores are aggregated to municipality, to ensure each geographical unit has a minimum amount of observations but the results are still meaningful for local interpretation. This aggregated score represents the subjective liveability per municipality, and is used as input for the comparison to the Leefbaarometer. Appendix A shows the code used for gathering the Twitter data. The code used for preparing the data for Hydrator, pre-processing the Hydrator output, and the performance assessment can be found in the supplementary files.

*Spatial analysis*

To find whether the sentiment scores show spatial patterns, a visual and a statistical analysis was done. The statistical analysis consists of performing a global and a local Moran's I, to test for clustering and clusters respectively. Moran's *I* is a spatial autocorrelation statistic, indicating whether there is a relation between neighbouring values or the values are randomly distributed. The global Moran's *I* indicates whether similar values are spatially concentrated or not over the complete research area and is visualised in a Moran scatter plot (Anselin et al., 2006). A Moran's *I* between 0 and 1 means similar values are close-by, and a Moran's *I* between -1 and 0 means opposite values are close-by. Figure 3.1 shows visualisations of both ends of the spectrum. The local Moran's I follows the same principles, but indicates clustering per geographical unit. This way, geographical areas where the clustering occurs can be identified, so called hotspots for  high value clusters and cold spots vice versa (Baele et al., 2010; Chi & Zhu, 2007). Spatial statistics like Moran's *I* are used to show patterns that are not immediately evident from non-spatial statistical tests, and quantify the spatial patterns that may emerge from the visual analysis (Baele et al., 2010).

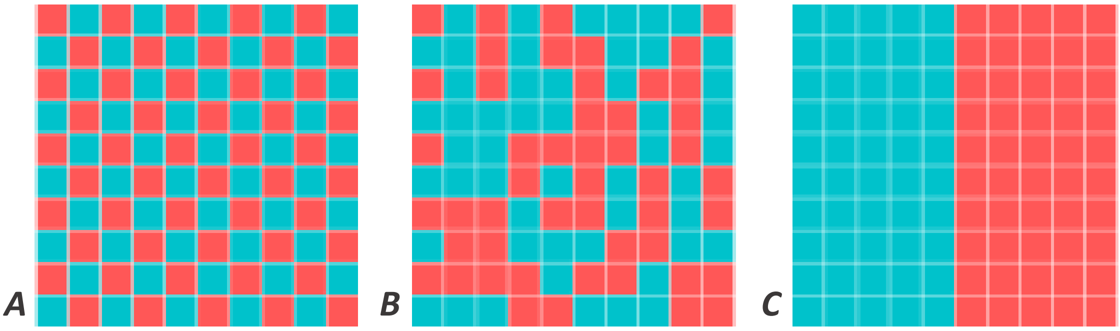


*Figure 3.1: Visualisation of Moran's I scores. In figure 3.1A, opposite values are close by, meaning the Moran's I will be close to -1. In figure3.1B, values are randomly distributed. The Moran's I will be close to 0. Figure 3.1C shows a distribution where similar values are close together, meaning the Moran's I will be close to 1.*

Whether values are considered geographically close or not in the calculation of the Moran's *I* depends on the spatial weights matrix used. A spatial weights matrix is an *n * n* positive symmetric matrix, as seen in equation 9. In this case, *n* represents the number of municipalities. Weight $w_{ij}$ describes the distance between location *i* and location *j*. If locations *ij* are neighbours according to the structure, $w_{ij}$ = 1, otherwise $w_{ij}$ = 1 (Suryowati et al., 2018).

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & w_{ij} & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \cdots & w_{nn} \end{bmatrix} \tag{9}$$

The spatial weights can be based on contiguity, geographical distance, or K-NN (Chi & Zhu, 2007). Contiguity weight matrices are based on the sharing of edges or corners between areal units. The geographical distance weight matrix is based on the theoretical distance at which the phenomenon is relevant, and considers all areal units with their centre within this theoretical distance neighbours. The K-NN weights are constructed to contain the *k* number of nearest neighbours specified. An issue directly related to using man-made areal units like municipalities in spatial analysis is the modifiable areal unit problem (MAUP). The MAUP occurs when the results of spatial analysis are influenced by scale and shape of the geographical unit used (Badland et al., 2014; Chi & Zhu, 2007). There are no generic solutions for this problem, but it should be taken into account in the discussion of results of spatial analyses (Badland et al., 2014; Openshaw & Taylor, 1979, as cited in Venerandi, 2017). Because the expected influence is in the direct area of the municipalities, a spatial weight matrix based on queen contiguity is used.

A common problem with contiguity weight matrices is the occurence of 'isolates', neighbourless areal units. The islands in the north of the Netherlands are not connected via land, they do not share a edge or corner with other municipalities. This means as far as the spatial weights matrix is concerned, they are isolated. However, in real life they are connected to the mainland by ferry services. To connect the islands in the north with the mainland, the ferry services between the corresponding municipalities were added manually as an extension of the municipality.

### 3.4 Comparison to Leefbaarometer

The comparison of the sentiment scores to the Leefbaarometer is done in three steps: The first step score is a visual analysis of maps depicting the Leefbaarometer classes and the sentiment scores. To allow for direct visual comparison, both maps are classified to classes of equal size. For the second and third step, the municipalities have been grouped to their corresponding Leefbaarometer classes. In step two, the mean sentiment scores of the three groups are compared in a graph. The third and last step statistically compares the population means. Because there are more than two independent populations, the statistical test used is an analysis of variance (ANOVA). The ANOVA tests whether the means of all populations are equal. If, based on the first two steps, the mean sentiment score looks like it increases with higher Leefbaarometer classes, and the ANOVA proves the means are not equal between the three populations, the sentiment scores can likely be used to represent liveability in the Netherlands.

# 4. Results

In this chapter, the most important outcomes of the performance analysis, sentiment analysis, and comparison to the Leefbaarometer are highlighted. The NB classifier with word n-gram (n = 1) using Dutch tweets showed the best performance in the test and was implemented in the sentiment analysis of the index dataset. The statistical distribution of the mean sentiment per municipality, or sentiment scores, can be assumed to be a normal distribution. The sentiment scores are more evenly spread over the research area than expected from a random distribution, and show significant local clustering in the more peripheral regions. A relation between sentiment scores and the Leefbaarometer classes seemed to emerge from a cursory analysis, but no statistical relation between the two was found.

## 4.1 Performance assessment

This section reports the results obtained from the performance assessment of the 8 different methods. The results of the 8 methods with the best performing parameters are shown in table 4.1, with the corresponding parameters in table 4.2. The NB algorithm with Dutch input performs the best on the test dataset, with a Youden's *J* of 0,46. The accuracy and $F_1$-score of the lexicon from De Smedt & Daelemans (2012a) are the highest of the tested classifiers, and based on the accuracy and $F_1$-score hybrid method of De Smedt & Daelemans (2012a) and K-NN also outperforms the NB, but the *J* of both is lower. This is because this classifier performs less when it comes to identifying negative tweets. The difference in performance in the classification of positive and negative tweets is remarkable. Using the most suitable parameters, all classifiers show a better true positive to false positive ratio than true negative to false negative ratio. The numbers and statistics accuracy, $F_1$-score, and Youden's *J* are based on can be found in appendix B, together with the statistics for the other tested parameters.

| **Category** | **Classifier** | **Classified (%)** | **Accuracy** | **$F_1$** | **Youden's *J*** |
|---|---|---|---|---|---|
| Lexicon-based methods | De Smedt & Daelemans (2012a) | 40 | 0,77 | 0,84 | 0,39 |
| | EmoWordNet | 97 | 0,51 | 0,48 | 0,22 |
| Supervised learning methods | NB Dutch | 100 | 0,68 | 0,71 | 0,46 |
| | NB English | 100 | 0,59 | 0,62 | 0,30 |
| | K-NN Dutch | 100 | 0,66 | 0,66 | 0,36 |
| | K-NN English | 100 | 0,68 | 0,78 | 0,33 |
| Hybrid methods | De Smedt & Daelemans (2012a) + NB | 99 | 0,72 | 0,77 | 0,39 |
| | De Smedt & Daelemans (2012a) + K-NN | 99 | 0,74 | 0,83 | 0,33 |
| | EmoWordNet + NB | 100 | 0,51 | 0,49 | 0,23 |
| | EmoWordNet + K-NN | 100 | 0,51 | 0,49 | 0,23 |

*Table 4.1: Performance of the tested classifiers.*

| Classifier | Threshold | Feature type |
|---|---|---|
| De Smedt & Daelemans (2012a) | 0,00 | n/a |
| EmoWordNet | 0,10 | n/a |
| NB Dutch | n/a | word n-gram (n = 1) |
| NB English | n/a | word n-gram (n = 1) |
| K-NN Dutch | n/a | word n-gram (n = 1) |
| K-NN English | n/a | word n-gram (n = 2) |
| De Smedt & Daelemans (2012a) + NB | 0,00 | word n-gram (n = 1) |
| De Smedt & Daelemans (2012a) + K-NN | 0,00 | character n-gram (n = 2) |
| EmoWordNet + NB | 0,10 | character n-gram (n = 1 or 3) |
| EmoWordNet + K-NN | 0,10 | word n-gram (n = 1) or character n-gram (n = 3) |

*Table 4.2: Optimal parameters for the tested classifiers.*

Based on the results of the performance assessment, the NB algorithm with word n-grams (n = 1) is chosen to be implemented in the sentiment analysis of the index dataset. Appendix C shows the code used to classify the index dataset. The results of this analysis are presented in the section below.

### 4.2 Sentiment analysis

The following section presents the outcomes of the sentiment analysis. First, some basic characteristics of the resulting dataset are presented. After this, the dataset is checked for normality of distribution and outliers. Last, the spatial patterns emerging from the sentiment scores are examined.

Using the trained NB classifier, the pre-processed tweets from the index dataset are analysed. In total, 1 375 203 tweets have been classified to 2 classes: positive or negative. 39% of the classified tweets was classified as positive and 61% as negative. The balance is similar to the results of the performance test, where the NB classifier predicted 43% of the tweets to be positive and 57% to be negative. Table 4.3 shows the exploratory statistics of mean tweet sentiment per municipality. The mean sentiment score per municipality is -0,221, and the number of classified tweets per municipality is shown in appendix C.

| Statistic | Value | Std. Error |
|---|---|---|
| Mean | -0,223 | 0,003 |
| Standard deviation | 0,058 | |
| Minimum | -0,576 | |
| Maximum | 0,111 | |
| Skewness | -0,616 | 0,127 |
| Kurtosis | 3,719 | 0,253 |

*Table 4.3: Exploratory statistics for tweet sentiment score per municipality.*

Ten of the 380 sentiment scores show values more than 3 standard deviations (0,074) from the mean sentiment score. These extreme scores are all compromised of less than 100 classified tweets, compared to a median value of 1251 tweets and a mean of 3487 tweets. The values have been treated as outliers and excluded from the non-spatial statistical tests. In appendix D a table with the values and corresponding municipalities can be found. After the removal of the outliers, the distribution of the sentiment scores was tested for normality. Table 4.3 shows the descriptive statistics for the sentiment scores excluding the outliers. Because the number of observations is greater than 300, the absolute skewness and kurtosis are used to determine normality instead of the Z-scores of these statistics (Kim, 2013). Because the skewness is between -2 and 2 and the kurtosis is below 4, normality can be assumed. The skewness of -0,616 shows a slightly negative skew compared to the perfect normal distribution, and the kurtosis of 3,719 means the outliers are more extreme than in the perfect normal distribution. The histogram in figure 4.1 visually confirms these patterns. Concluding, the distribution of sentiment scores is assumed to be normal.

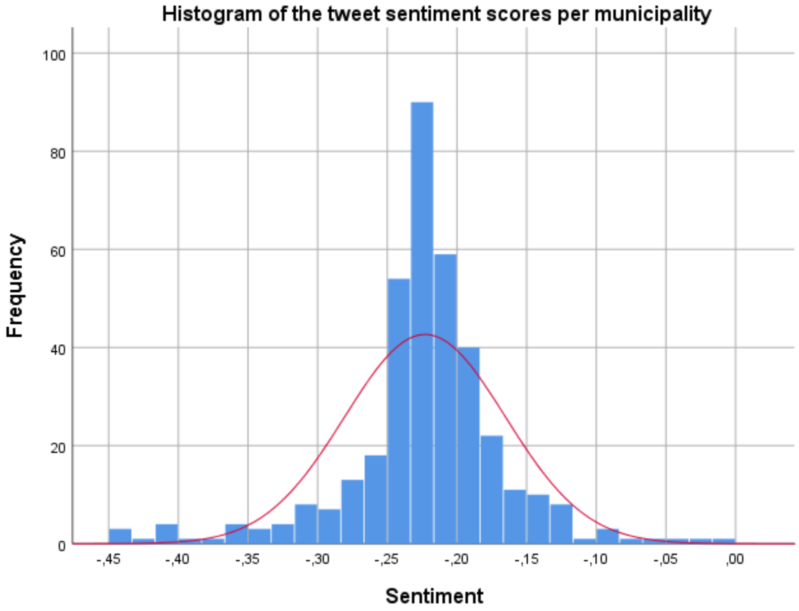


*Figure 4.1: Frequency distribution of tweet sentiment scores per municipality.*

Figure 4.2 shows the mean tweet sentiment per municipality in the Netherlands. Globally, the more extreme values, positive as well as negative, occur more often in the peripheral regions. The western, more urbanized region shows more moderate values. Regarding local patterns, a few can be spotted. In the South of Zeeland, the East of Overijssel and the North of Friesland and Groningen, four clusters of higher values can be seen. A small group of lower values can be seen in the North of Utrecht and West of North Holland, and in the South of Overijssel.

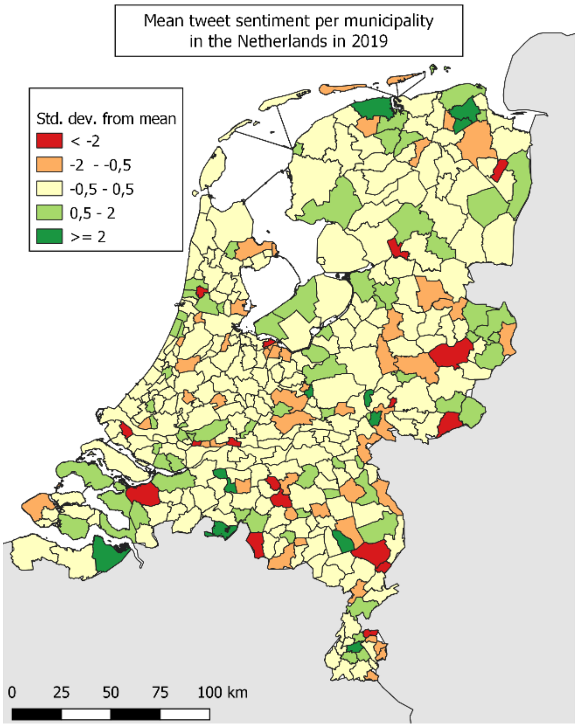


*Figure 4.2: Mean tweet sentiment per municipality in the Netherlands in 2019.*

Global and local spatial autocorrelation is tested using Moran's *I*. The spatial weight matrix used in both the global and the local Moran's *I* test is based on a second degree queen contiguity. Based on the visual analysis of figure 4.2, the global autocorrelation is expected to be insignificant, as the extreme values do seem to occur close to each other, but not ordered. The null hypothesis for the global Moran's *I* is that the values are randomly dispersed. The alternative hypothesis is that similar values are spatially close-by. The Moran's *I* is -0,036, indicating a small negative spatial autocorrelation. This means low values are more likely to be close to high values and vice versa. Based on 999 permutations, the null hypothesis is rejected (pseudo-p = 0,025, z = -1,809). This means the negative global spatial autocorrelation is significant at a confidence level of 0,05. The z-value is negative, meaning the high and low values are more spatially spread out than expected in a random distribution.

The local spatial autocorrelation is expected to show clusters in the areas mentioned in the visual analysis. High-high clusters are expected in the South of Zeeland, the North of Groningen, the North of Friesland and the East of Overijssel. Two small low-low clusters are expected in in the North of Utrecht and the South of Overijssel. The null hypothesis is that the municipality is surrounded by random values, with an alternative hypothesis that the municipality is surrounded by either opposite values or similar values. Figure 4.3 shows the clusters that emerged from the local Moran's *I*. The high-high clusters from the North of Groningen and South of Zeeland that emerged from the visual analysis are significant, but the other visual patterns are not confirmed by the local Moran's *I*. The outcomes of the Queen's contiguity order comparison, the Moran's *I* scatterplot, and the significance of the sentiment score clusters in figure 4.4 can be found in appendix E.

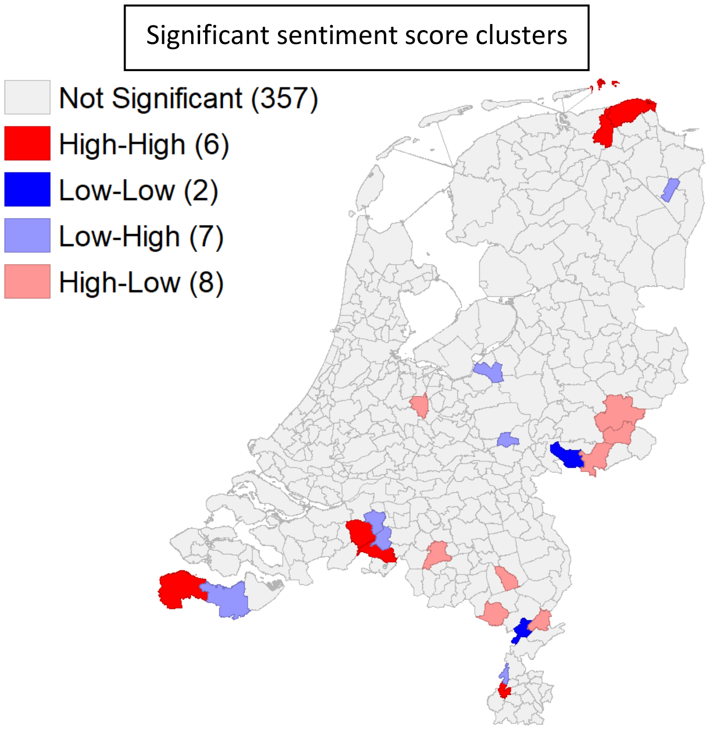


*Figure 4.4: Types of clusters from the local Moran's I of sentiment scores.*

Concluding, the distribution of sentiment scores is approximately normal, after excluding the ten municipalities that show extreme values. Globally, a weak pattern emerged from the visual analysis, and a weak, statistically significant pattern was found in the spatial distribution sentiment scores. Locally, a few clusters emerged from both the visual and statistical analysis.

### 4.3 Comparison to Leefbaarometer

This section focusses on comparing the sentiment scores with the Leefbaarometer 2.0, version 2018. As mentioned in section 3.4, the Leefbaarometer index consists of ordinal data: the results are grouped to ordered classes. The 9 original classes have been grouped to 3 classes for this research. These three remaining classes are used as the grouping variable for the municipalities in the comparison to the sentiment scores. The comparison of the Leefbaarometer to the sentiment scores is done in three steps: a visual analysis of spatial patterns, a visual comparison of the group means, and a statistical comparison of the group means.

Figures 4.5 and 4.6 show the reclassified Leefbaarometer scores and the reclassified sentiment scores respectively. Drenthe, Overijssel, and the East of Utrecht show predominantly high Leefbaarometer scores, with Flevoland, Zeeland, the South of South Holland, and the East of Groningen showing mostly low scores. Most of these patterns are not replicated in the sentiment scores. In fact, one of the patterns is reversed: Zeeland and the South of South Holland show mostly high sentiment scores. The high sentiment scores in the South of Overijssel are in accordance with the Leefbaarometer scores, as well as the small group of high values surrounded by lower values in the South of Limburg. Visually, there is little spatial similarity between the Leefbaarometer and the sentiment scores.

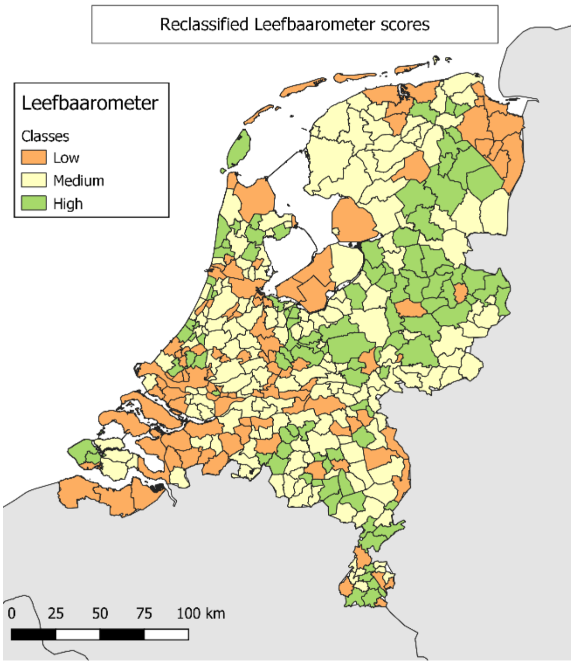


*Figure 4.5: Reclassified Leefbaarometer 2.0, version 2018.*

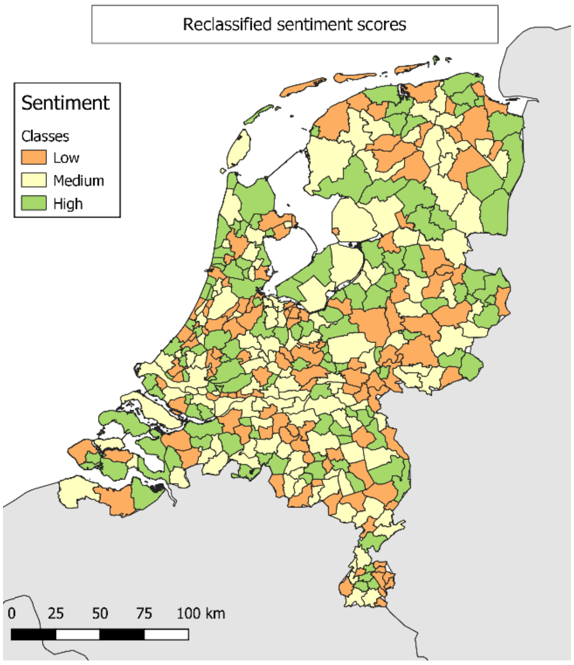


*Figure 4.6: Reclassified sentiment scores per municipality in the Netherlands in 2019.*

For the visual comparison of the group means, figure 4.7 is used. The group of municipalities with a low or medium Leefbaarometer class show a mean sentiment score below the overall mean. The group of municipalities with a high Leefbaarometer score show a population above the overall mean. The difference between the low and medium group is relatively small compared to the difference between the high group and the other two. Following this graph, the means are expected to be significantly different in the ANOVA.

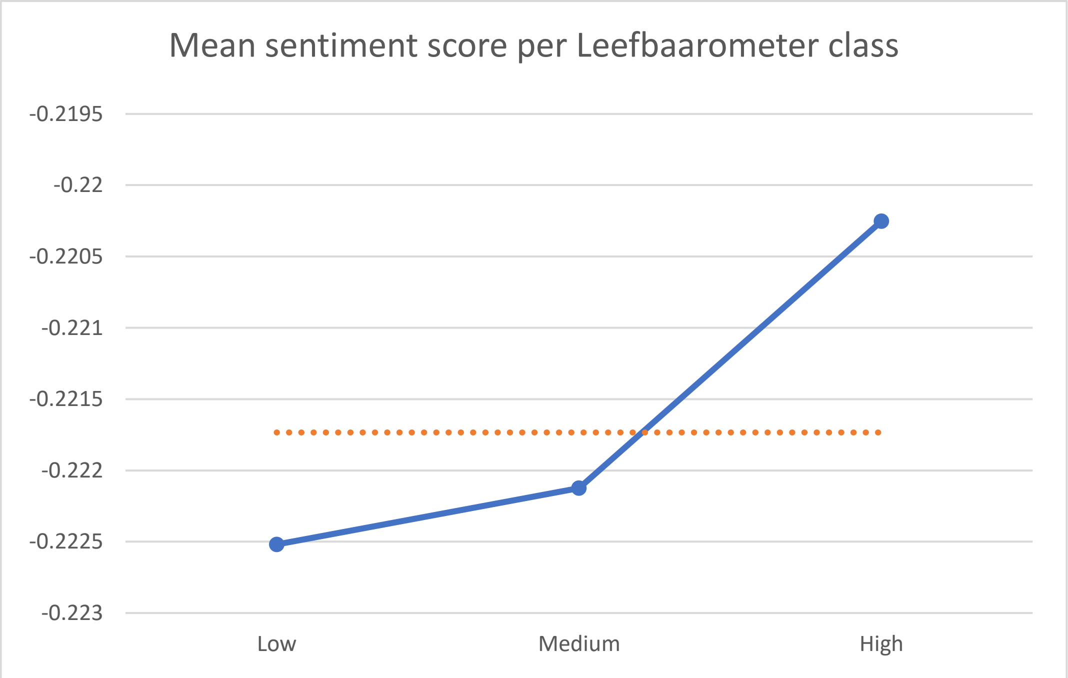


*Figure 4.7: Mean sentiment score per Leefbaarometer class in the Netherlands in 2019.*

*Figure 4.7: Mean sentiment score per Leefbaarometer class*

The three populations of sentiment scores meet the prerequisites of normal distribution and equal group variances. Based on the null hypothesis that the means of the populations are equal, and an alternative hypothesis that the means vary significantly between populations. The results of the ANOVA show that the at a significance level of 0,05 the means do not vary between the groups (F = 0,035; p = 0,462) and the null hypothesis is accepted. This means the sentiment scores do not significantly differ between different Leefbaarometer classes, rejecting the hypothesis formed in the previous paragraph. This leads to the conclusion that the sentiment scores show no significant similarities to the Leefbaarometer.

# 5. Discussion

The concept of liveability is hard to assess and quantify because of the ambiguity and complexity of the definition and the contributing factors. Flaes et al. (2016) and Mitchell et al. (2013) relate sentiment and emotions of residents to aspects of liveability, but the use of sentiment as a direct indicator has not yet been studied. The gap in liveability between the urbanised West and the more peripheral regions has widened over the past years in the Netherlands, but the local liveability differences in the peripheral regions have also grown. This indicates a need for both national and regional liveability assessments. In the Netherlands, the relation between sentiment and several aspects of liveability has been studied at city level, but not for the Netherlands as a whole.

Various approaches for sentiment analysis of Twitter content have been studied before(e.g. Agarwal et al., 2011; Huq et al., 2017; Jianqiang & Xiaolin, 2017), but mostly for tweets from countries where English is the primary language, or where all non-English tweets were filtered out. Because of this, premade classifiers for English tweets are widely available (Madhoushi et al., 2015). For Twitter content from countries like the Netherlands, little research on the best practices regarding sentiment analysis is available.

Following this, the main goal of this study is to evaluate the relation between sentiment and liveability in the Netherlands. The secondary, more methodological objective is to explore different approaches of extracting sentiment from UGGC in the Netherlands.

## 5.1 Key findings

The results of this study showed no relation between sentiment derived from UGGC and liveability in the Netherlands. Thus, sentiment cannot be used as a direct indicator of liveability on a national scale in the Netherlands, However, it can be used to identify regions with significant local differences in liveability because the expected regional liveability discrepancies were also found in the spatial analysis. The methodological objective showed that based on Youden's *J* and percentage of tweets classified, the approach best suited for extracting sentiment from UGGC in the Netherlands translates all content to Dutch, groups the text to word n-grams with n = 1, and classifies them using a NB algorithm.

The best approach to classify UGGC from the Netherlands to sentiment is to translate all tweets to Dutch and use a learning-based sentiment analysis method, specifically a NB algorithm, with a Youden's *J* of 0,46 and 100% of the tweets classified. This answers the first research question. All classifiers performed better on the dataset where less of the tweets had to be translated. This means that translating the text to a different language reduces the performance, in this study the Youden's *J* decreased by 0,10 when 15% more tweets had to be translated. The learning-based methods also all performed the best using word n-grams, specifically using lower values for n. This means words that often occur close to each other in a sentence have similar sentiment values, more so than words that do often occur in the same context but do so further apart from each other.

The second key finding of this study is that sentiment in the Netherlands is more evenly spread compared to the expected liveability distribution, and does not show the expected higher scores in the West compared to the rest of the Netherlands. The expected regional liveability differences in peripheral regions were present in the sentiment scores, as most of the significantly conflicting clusters from the local Moran's *I* are located there. These conclusions mean that sentiment derived from UGGC cannot be used to assess liveability at a national scale in the Netherlands, but can be used to identify regions with significant local differences in liveability.

A weak visual relation between the Leefbaarometer and the sentiment scores was observed, where the mean sentiment score seemed to increase with higher Leefbaarometer classes. This relation is not confirmed by the ANOVA, thus no relation could be proven between the sentiment scores and the liveability index. This difference could come from the difference of interpretation of the Leefbaarometer classes. In the visual analysis, the classes are interpreted as ordered, whereas the ANOVA interprets them as categorical. Secondarily, the means of the classes lie within 0,003 of each other. This is a small difference compared to the total range of sentiment scores, which is from -0,443 and 0,000. The absence of a statistical relation confirms the conclusion from the previous paragraph that sentiment cannot be used to assess liveability at a national scale in the Netherlands.

## 5.2 Other research

To assess in what ways this study adds to the existing discourse on sentiment extraction from UGGC and the relation between sentiment and liveability, a comparison is made between similar academic research and the outcomes of this study. The findings that learning-based methods perform the best on Dutch tweets conflicts with the work from Flaes et al. (2016). The negative effect of translation on the performance of the classifier mirrors the results in Basiri & Kabiri (2017) and in Manad et al. (2019). The performance of the NB classifier with a small training dataset concur with the results of Abdelwahab et al. (2016). The relation between aspects of liveability and sentiment found in Flaes et al. (2016) and Mitchell et al. (2013), did not translate to a relation between sentiment and liveability as a whole.

The approaches solely using learning-based methods performed better compared to approaches using premade sentiment classifiers in this study. In research using content from English-speaking countries, premade classifiers performed better than the learning-based classifiers (Badaro et al., 2018; Fathy et al., 2017). Premade classifiers from countries with a smaller primary language, like the lexicon from De Smedt and Daelemans (2012a), are likely not as far in their development as the English counterparts. The Dutch lexicon was built from the base and published in 2012 and has not seen further development, whereas EWN was published in 2018, but was built on DepecheMood, an emotion lexicon published by Staiano and Guerini (2014), and used components from multiple other lexical resources (Badaro et al., 2018; De Smedt & Daelemans, 2012a). The finding that learning-based classifiers performs the best conflicts with comparative research using English content (e.g. Agarwal et al., 2017; Jianqiang & Xiaolin, 2017) and research using content in other languages (e.g. Flaes et al., 2016; Rogers et al., 2018), even though the performance of the lexicon-based methods is comparable to the results in the studies they were proposed in (Badaro et al., 2018; De Smedt & Daelemans, 2012a). Based on the common performance measures $F_1$-score and accuracy, the lexicon-based method would have been the most effective too. The use of a different selection criterion can explain the discrepancy between previous research and this study.
Compared to the performance of NB classifiers in other sentiment analysis research, the classifier used this study underperformed. For instance, in Go et al. (2009b) the NB classifier showed an accuracy of 84%, compared to the 68% in this study. The underperformance can largely be attributed to the size of the training dataset: Go et al. (2009b) used 10 000 to 150 000 tweets to train their classifiers, where the classifiers in this study were trained and tested using 570 and 235 tweets respectively. For a NB classifier with a training dataset consisting of 570 tweets, the performance was as expected when compared to the results of a NB classifier trained with 470 tweets in Abdelwahab et al. (2016). Concluding, the performance of the NB classifier in this study is similar to NB classifiers in research with similar training datasets.

The relevance of sentiment in the identification of regional liveability discrepancies shown in this study has not been shown in previous research. Other research on liveability and sentiment derived from UGGC either focussed on contrasts within cities (e.g. Flaes et al., 2016; Kovács-Győri et al., 2018), or used a larger areal unit like states or provinces (e.g. Javadi & Taleai, 2020; Mitchell et al., 2013). Studies that did investigate liveability on a regional scale, used more traditional methods to gather their data, like surveys or in-depth interviews (e.g. Beunen et al., 2020; Gieling & Haartsen, 2017; Ubels et al., 2019). Because of the growing popularity of LBSNs, the use of UGGC in research in less densely populated areas is becoming more feasible.

Contrary to the hypothesis based on previous research, no statistical relations were found between liveability and sentiment in the Netherlands in this study. Mitchell et al. (2013) concluded that sentiment derived from Twitter messages correlated with wealth and anti-correlated with obesity, two contributing factors of liveability, and Flaes et al. (2016) added a correlation between tweet sentiment and education level to this. These correlations to aspects of liveability in other research do not translate to a statistical relation between sentiment and liveability in this study. These conflicting outcomes can be explained by the idea that liveability as a whole is a complex concept that is hard to assess and quantify, as previously stated. The use of one aspect clearly does not cover the complexity of liveability. Another possible explanation for the disparity of results between existing liveability research and this study, is the scale effect. When the same data is analysed over different scales, the results can differ (Chi & Zhu, 2008). The research by Mitchell et al. (2013), Flaes et al. (2016), and this study all use different areal units. Flaes et al. (2016) analysed the relation between liveability and UGGC derived sentiment within a city and Mitchell et al. (2013) at state level and between different cities, whereas this study used municipalities as the areal unit.

### 5.3 Limitations and opportunities

There are different aspects of this study that introduce limitations to the generalisability of the findings of this study, but also aspects that provide opportunities for future research. These limitations and opportunities need to be taken into account when applying the findings of this study to a different context. The limitations and opportunities are grouped to seven categories.

The Twitter data used was assumed to represent the whole population of the Netherlands in 2019. This is likely not true, as the characteristics of the users of Twitter do not exactly mirror the characteristics of all residents (Kovács-Győri et al., 2018; Zivanovic et al., 2020). Age and social-economic situation are some of the factors that influence liveability, but are likely not correctly represented in the data used (Badland & Pearce, 2019; Badland et al., 2014; Leidelmeijer & van Kamp, 2003; Namazi-Rad et al., 2012). Both very young and very old populations are expected to be underrepresented, as well as poorer populations (Kovács-Győri et al., 2018). Wealthy and relatively young people are likely overrepresented in the data (Zivanovic et al., 2020). To help limit the impact of the characteristics of twitter users compared to the characteristics of the general population of the research area, the weight of the sentiment scores of users can be adjusted. This could be done by collecting some basic characteristics of the users of the LBSN the UGGC is collected from, and correcting the proportions to mirror the proportions of the population of the research area. This helps the generalisability of the findings, but was outside the scope of this study.

Another disadvantage of using UGGC and a sentiment analysis to measure emotions of residents, is that the spontaneously generated content does not necessarily represent expression of emotions or satisfaction with the environment (Kovács-Győri et al., 2018; Nenko & Petrova, 2019). It can be seen as a form of revealed preference however, justifying the assumption that it can be used as an indicator of emotion or satisfaction (Tieskens et al., 2018). A more in depth look into the relation between sentiment and liveability at an individual level could provide further insight into the exact place of sentiment in regional or national liveability research.

Some tweets contain exact coordinates, but the majority is geotagged to often a well-known location. By including the geo-tagged tweets, the location of the sentiment score could be different from the place where the sentiment was generated. If only the tweets with exact coordinates are used, the location of the sentiment score is exactly the place where this sentiment was generated. However, this would have resulted in losing 91% of the collected tweets and thus led to a strong selection of data, limiting its generalisability. If this choice is made nonetheless, the aforementioned correction for population characteristics becomes even more important because of the stricter selection.

By choosing municipalities as the areal unit to which the sentiment was generalised, the MAUP was introduced. The shape and scale of areal units influence the outcomes of analyses (Badland et al., 2014; Chi & Zhu, 2007). The spatial weights matrix used in the spatial autocorrelation analysis was based on the sharing of borders, enhancing the influence of the shape of the municipality on the spatial autocorrelation analysis. The patterns shown in the analysis may differ when the spatial autocorrelation of sentiment is investigated at another scale, for instance at neighbourhood-level or provincial level, or at another type of areal unit, for instance by dividing the research area in evenly sized squares. This areal unit was chosen based on two factors. The first factor is because most of the geo-tags of the twitter data are population centres, making an analysis on a more precise level than cities not feasible. The second factor was in which areal units the existing liveability index was available, of which municipalities were the closest scale to population centres.

The method chosen for the detection of automated users or bots is based on the behaviour of the user, but is not waterproof. The research in which this method is proposed mentions that it will not detect all automated users, but by filtering out the users that post the most tweets, the most impactful automated users are filtered out (Shao et al., 2018). However, this does mean that the non-automated users that tweeted more than 15 times a day were also filtered out. The automated users that tweet less than 15 times a day were not filtered out, but their impact on the outcomes is low because of the relatively low volume of tweets.

The textual anomalies such as slang or words with extra repeated vowels were not adjusted to normal words in this study. This has likely negatively impacted the performance of the lexicon-based classifiers in the performance assessment, possibly also the performance of the other classifiers (Agarwal et al., 2011; Baid et al., 2017; Giachanou & Crestani, 2016; Jianqiang & Xiaolin, 2017). Because difference between the best performing classifier and the other classifiers in Youden's *J* was relatively large, the likelihood that this omission impacted the final choice of classifier is low.

The messages of the tweets in both datasets were manually classified by the author, possibly introducing a personal bias. The amount of training and test data was also limited due to the manual classification of one person. The choice for manual classification was made, because for English tweets more extensive resources of classified tweets are available, but for Dutch tweets no pre-classified twitter data was found. The influence of personal bias can be decreased either by involving more people in the manual classification, or reaching for external pre-classified datasets. The use of external pre-classified datasets likely also increases the size of the training and test datasets, which in turn increases the performance of the learning-based classifiers. Constructing and publishing pre-classified datasets for different languages in future research will improve the quality of sentiment analysis in non-English speaking countries.

This study compared two lexicons, two learning-based methods, and four hybrid methods in the performance assessment. The NB classifier performed the best out of these, but there are more lexicon-based and learning-based methods available, as well as more ways to combine these two into hybrid methods. This means the answer to the first sub-question is not a definitive answer to what the best sentiment classifier is, but a good baseline for more exploration on different classification methods. Including other learning-based classifiers, like support vector machine, or unsupervised learning-based classifiers, like long short-term memory and convolutional neural network approaches, in a future performance assessment for sentiment classification can be a valuable extension.

The choice of Youden's *J* as the primary performance measure impacted the choice of classifier. In other studies, $F_1$ scores are commonly used to determine which sentiment classifier performs the best. By choosing Youden's *J*, performance in identifying negative and positive tweets is equally important, where $F_1$ scores emphasize performance in identifying positive tweets. This is confirmed by the performance of de hybrid method combining De Smedt & Daelemans (2012a) and K-NN: based on percentage classified and $F_1$ score, this would have been the best classifier, but misclassified more than 70% of the negative tweets. This insight should be taken into account in future research where classifiers are compared.

One of the obstacles in the spatial analysis was the absence of shared borders between the islands in the Wadden sea in the North of the Netherlands and the mainland. This issue was resolved by manually connecting the islands to the mainland where the ferry services connect them in reality. This does mean that the connection between them is assumed to be the same as between two municipalities that are direct neighbours. The travel time and monetary fee that the ferry connection introduces are likely to impair the actual connectivity of these municipalities. This likely reduces the strength of the liveability relation between the islands and the mainland municipalities. In future research, the ferry connection can be weighted to correct for this.

The lack of research on the relation between sentiment derived from UGGC and liveability at municipal level presents an opportunity for future research. By using UGGC to analyse the possible regional discrepancies instead of traditional methods, temporal and monetary costs can be reduced. After the indication of these discrepancies, more in depth research can be conducted to the underlying factors and possible solutions to improve liveability in the areas that are lagging behind.

The conclusion that UGGC derived sentiment can be used to identify regional differences in liveability is limited by the areal unit used: the conclusion is valid for analysis at municipal level. Research using states, provinces, or neighbourhoods within a city came to different conclusions, implying that there might be a significant influence of spatial scale on the relation. An opportunity for future research is to evaluate the influence of scale or areal units on the relation between sentiment derived from UGGC and liveability.

The Leefbaarometer is offered in a ordinal data format, limiting the choices for statistical analysis. The Leefbaarometer consists of nine classes, of which six are populated in the version of 2018. Of these six, the lowest two and the highest consist of too little cases to provide insight into possible relations with sentiment. This means that the Leefbaarometer was reduced to three classes, but these three classes all were sufficiently filled for statistical analysis. However, the reclassification decreased the precision of the data, limiting the precision of the analysis.

The visual analysis of the relation between sentiment and the Leefbaarometer concluded that a weak trend seemed to exist. This cursory analysis interpreted the Leefbaarometer classes as ordered. For the non-spatial analysis, an ANOVA was chosen with the Leefbaarometer as the grouping variable, and analysed whether one class mean differed significantly from the means of the other classes, disregarding the possible trend, and found no relation. By choosing ANOVA as the statistical test, the normality and ratio format of the sentiment scores were kept intact, but the ordered format of the Leefbaarometer classes was dropped. By using a liveability index in either ratio or interval data format, more spatial and non-spatial statistical tests can be used, opening up the option for trend testing.

## 6. Conclusion

Residents' sentiment is closely associated with liveability, and when compared to contributing factors to liveability or included as an independent variable in a liveability model, significant relations were shown. This study explored the relation of the sentiment to liveability as a whole at a municipal level. Based on a literature review, the sentiment was expected to show similarities to liveability. This was tested by gathering the residents' sentiment from twitter messages, and comparing the mean sentiment to an expected liveability discrepancy pattern and an existing liveability index. The twitter-derived sentiment showed little similarities to both the expected patterns and the existing liveability index. In the peripheral regions, the expected larger differences did show in the residents' sentiment. Also visually, a trend between the Leefbaarometer groups and the mean sentiment seemed to exist, however no statistical relation  was found.

The mean sentiment in municipalities showed some similarities to liveability in cursory analysis, but no statistical relations were found, leading to the conclusion that the sentiment derived from twitter messages in this study does not significantly represent liveability.

In addition, this study builds on previous research by exploring the possibilities of including UGGC in liveability research, building on previous research. The learning-based classifiers used for sentiment analysis in this study show promising results, even with relatively small training and test datasets. Exploring the performance of other classifiers or implementing the methods used in this study using more training and test data could further solidify the conclusive strength of research regarding twitter-derived sentiment and liveability. Even though this study did not show a correlation between residents' sentiment from twitter messages and liveability, it does not mean there is no relation between sentiment and parts of liveability. Dissecting liveability and comparing its different aspects to the sentiment of residents using various areal units could be another valuable extension of this study.

## 7. References

Abdelwahab, O., Bahgat, M., Lowrance, C. J., & Elmaghraby, A. (2015). Effect of training set size on SVM and Naïve Bayes for Twitter sentiment analysis. *2015 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2015,* 46-51. https://doi.org/10.1109/ISSPIT.2015.7394379

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter data. *Proceedings of the Workshop on Language in Social Media*, 30-38.

Alfarrarjeh, A., Agrawal, S., Kim, S. H., & Shahabi, C. (2017). Geo-spatial multimedia sentiment analysis in disasters. Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017, 2018-Janua, 193–202. https://doi.org/10.1109/DSAA.2017.77

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis, 38*(1), 5-22. https://doi.org/10.1111/j.0016-7363.2005.00671.x

Antognelli, S., & Vizzari, M. (2017). Landscape liveability spatial assessment integrating ecosystem and urban services with their perceived importance by stakeholders. *Ecological Indicators, 72*, 073-725. http://doi.org/10.1016/j.ecolind.2016.08.015

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert systems with Applications, 164*, 114006. https://doi.org/10.1016/j.eswa.2020.114006

Badaro, G., Jundi, H., Hajj, H., & El-Hajj, W. (2018). EmoWordNet: Automatic Expansion of Emotion Lexicon Using English WordNet. 86–93. https://doi.org/10.18653/v1/s18-2009

Badland, H., & Pearce, J. (2019). Liveable for whom? Prospects of urban liveability to address health inequities. *Social Science & Medicine, 232*, 94-105. https://doi.org/10.1016/j.socscimed.2019.05.001

Badland, H., Whitzman, C., Lowe, M., Davern, M., Aye, L., Butterworth, I., Hes, D., & Giles-Corti, B. (2014). Urban liveability: Emerging lessons from Australia for exploring the potential for indicators to measure the social determinants of health. *Social Science and Medicine, 111*, 64-73. http://dx.doi.org/10.1016/j.socscimed.2014.04.003

Baele, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., & Elston, D. A. (2010). Regression analysis of spatial data. *Ecology Letters, 13*, 246-264. https://doi.org/10.1111/j.1461-0248.2009.01422.x

Baid, P., Gupta, A., & Chaplot, N (2017). Sentiment Analysis of Movie Reviews using Machine Learning Techniques. *International Journal of Computer Applications, 179*(7), 45-49. https://doi.org/10.5120/ijca2017916005

Basiri, E., & Kabiri, A. (2017). Translation is not enough: Comparing Lexicon-based methods for sentiment analysis in Persian. *18th CSI International Symposium on Computer Science and Software Engineering,* 36-41. https://doi.org/10.1109/CSICSSE.2017.8320114

Beattie, L., & Haarhoff, E. (2018). Urban Growth, Liveability and Quality Urban Design: Questions about the efficacy of urban planning systems in Auckland, New Zealand. *Journal of Contemporary Urban Affairs, 2*(2), 12-23. https://doi.org/10.25034/ijcua.2018.3667

Beunen, R., Meijer, M., & de Vries, J. (2020). Planning strategies for dealing with population decline: Experiences from the Netherlands. Land Use Policy, 93. https://doi.org/10.1016/j.landusepol.2019.104107

Caron, J., Cargo, M., Daniel, M., & Liu, A. (2019). Predictors of Quality of Life in Montreal, Canada: A Longitudinal Study. Community Mental Health Journal, 55(2), 189–201. https://doi.org/10.1007/s10597-018-0340-y

Chi, G., & Zhu, J. (2008). Spatial Regression Models for Demographic Analysis. *Population Research and Policy Review, 27*(1), 17-42. https://doi.org/10.1007/s11113-007-9051-8

Clarke, A., & Cheshire, L. (2018). The post-political state? The role of administrative reform in managing tensions between urban growth and liveability in Brisbane, Australia. *Urban Studies, 55*(16), 3545-3562. https://doi.org/10.1177/0042098017753096

Conger, B. W. (2015). On Livability, Liveability and the Limited Utility of Quality-of-Life Rankings. *The School of Public Policy Publications, 7*(4), 1-9. https://doi.org/10.11575/sppp.v8i0.42528

Curini, L., Iacus, S., & Canova, L. (2015). Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case. *Social Indicators Research, 121*(2), 525-542. https://doi.org/10.1007/s11205-014-0646-2

Danielaini, T. T., Maheshwari, B., & Hagare, D. (2019). Qualitative and quantitative analysis of perceived liveability in the context of socio-ecohydrology: Evidence from the urban and per-urban Cirebon-Indonesia. *Journal of Environmental Planning and Management, 62*(12), 2026-2054. https://doi.org/10.1080/09640568.2018.1524576

Dave, A., Bharti, S., Patel, S., & Mishra, S. K. (2021). A Real-Time Sentiments Analysis System Using Twitter Data. *Smart Innovation, Systems and Technologies, 153*, 251-258. https://doi.org/10.1007/978-981-15-6202-0_25

De Smedt, T., & Daelemans, W. (2012a). “vreselijk mooi!” (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. In Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012 (pp. 3568–3572). European Language Resources Association (ELRA).

De Smedt. T., & Daelemans, W. (2012b). Pattern for python. *Journal of Machine Learning Research, 13,* 2063-2067.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. In Procedia Computer Science (Vol. 87, pp. 44–49). Elsevier B.V. https://doi.org/10.1016/j.procs.2016.05.124

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naïve Bayes’ and K-NN Classifier. *International Journal of Information Engineering and Electronic Business, 8*(4), 54-62. https://doi.org/10.1007/s10664-017-9546-9

Dredze, M., Paul, M. J., Bergsma, S., & Tran, H. (2013, July). Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI), 23*, 45-50.

Fathy, S., El-Haggar, N., & Haggag, M. H. (2016). A Hybrid Model for Emotion Detection from Text. *International Journal of Information Retrieval Research, 7*(1), 32–48. https://doi.org/10.4018/ijirr.2017010103

Fellbaum, C. (2006). Wordnet(s). In Brown, K (eds.), *Encyclopedia of Language & Linguistics, Second Edition* (vol. 13), 665-670. Oxford: Elsevier.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The Rise of Social Bots. *Communications of the ACM, 59*(7), 96-104. https://doi.org/10.1145/2818717

Flaes, J. B., Rudinac, S., & Worring, M. (2016). What multimedia sentiment analysis says about city liveability. In *European Conference on Information Retrieval*, 824-829. https://doi.org/10.1007/978-3-319-30671-1_74

Fu, B., Yu, D., Zhang, Y. (2019). The livable urban landscape: GIS and remote sensing extracted land use assessment for urban livability in Changchun Proper, China. *Land Use Policy, 87*. https://doi.org/10.1016/j.landusepol.2019.104048

Gieling, J., & Haartsen, T. (2016). Liveable Villages: The Relationship between Volunteering and Liveability in the Perceptions of Rural Residents. *Sociologia Ruralis, 57*(S1), 576-597. https://doi.org/10.1111/soru.12151

Ghiassi, M., Skinner. J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using *n*-gram analysis and dynamic artificial neural network. *Expert Systems with Applications, 40*(16), 6266-6282. https://doi.org/10.1016/j.eswa.2013.05.057

Giachanou, A. & Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys, 49*(2), 28. https://doi.org/10.1145/2938640

Go, A., Bhayani, R., & Huang, L. (2009a). Twitter Sentiment Classification using Distant Supervision. *Processing*, 1-6.

Go, A., Huang, L., & Bhayani, R. (2009b). Twitter sentiment analysis. *Entropy, 17,* 252.

Ho, H. C., Man, H. Y., Wong, M. S., Shi, Y., Walker, B. B. (2020). Perceived differences in the (re)production of environmental deprivation between sub-populations: A study combining citizens' perceptions with remote-sensed and administrative data. *Building and Environment, 174*, 1-14. https://doi.org/10.1016/j.buildenv.2020.106769

Hollander, J. B. (2011). Can a city successfully shrink? Evidence from survey data on neighborhood quality. Urban *Affairs Review, 47*(1), 129-141. https://doi.org/10.1177/1078087410379099

Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment Analysis on Twitter Data using K-NN and SVM. *International Journal of Advanced Computer Science and Applications, 8*(6), 19-25.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

Javadi, G., & Taleai, M. (2020). Integration of User Generated Geo-contents and Official Data to Assess Quality of Life in Intra-national Level. In Social Indicators Research (Issue 1346). Springer Netherlands. https://doi.org/10.1007/s11205-020-02437-1

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access, 5, 2870–2879. https://doi.org/10.1109/ACCESS.2017.2672677

Kapočiūtė-Dzikienė, J., Damaševičius, R., & Woźniak, M. (2019). Sentiment analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches. *Computers, 8*(1), 4. https://doi.org/10.3390/computers8010004

Kashef, M. (2016). Urban livability across disciplinary and professional boundaries. Frontiers of Architectural Research, 5(2), 239–253. https://doi.org/10.1016/j.foar.2016.03.003

Khomenko, S., Nieuwenhuijsen, M., Ambròs, A., & Wegener, S. (2020). Is a liveable city a healthy city? Health impacts of urban and transport planning in Vienna, Austria. *Environmental Research, 183*, 109238. https://doi.org/10.1016/j.envres.2020.109238

Kim, H-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution using skewness and kurtosis. *Restorative Dentistry & Endodontics, 38*(1), 52-54. https://doi.org/10.5395/rde.2013.38.1.52

Kotkin, J. (2009, August 10). *Why The 'Livable Cities' Rankings Are Wrong.* Forbes. https://www.forbes.com/2009/08/10/cities-livable-elite-economist-monocle-rankings-opinions-columnists-joel-kotkin.html?sh=67a4a6432181

Kovács-Győri, A. (2019). GIS-based Livability Assessment: A practical tool, a promising solution? GISTAM 2019 - Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management, Gistam, 289–296. https://doi.org/10.5220/0007753702890296

Kovács-Győri, A., & Reinel, B. (2017). Reflecting Individual Preferences and Spatiality in Livability Measurements: A Livability Assessment Platform for the City of Salzburg. In *E. Tracada & G. Cairns (Eds.), AMPS Proceedings Series 10 - Cities, Communities and Homes: Is the Urban Future Livable?*, 211–221. Derby, UK: AMPS C.I.O

Lapointe, M., Cumming, G. S., & Gurney, G. G. (2019). Comparing Ecosystem Service Preferences between Urban and Rural Dwellers. *BioScience, 69*(2), 108-116. https://doi.org/10.1093/biosci/biy151

Leidelmeijer, K., van Kamp, I. (2003). Kwaliteit van de Leefomgeving en Leefbaarheid: naar een begrippenkader en conceptuele inkadering. *RIVM rapport 630950002/2003.* Retrieved from https://www.rivm.nl/bibliotheek/rapporten/630950002.pdf

Leidelmeijer, K., Marlet, G., Ponds, R., Schulenberg, R., Van Woerkens, C., & Van Ham, M. (2014). Leefbaarometer 2.0: Instrumentontwikkeling. In Research en Advies, Atlas voor gemeenten. Retrieved from https://doc.leefbaarometer.nl/resources/Leefbaarometer%202.0%20Instrumentontwikkeling%20CONCEPT.pdf

Leidelmeijer, K., Middeldorp, M., & Marlet, G. (2019). Leefbaarheid in Nederland 2018: een analyse op basis van de leefbaarometer 2018. Retrieved from: https://leefbaarometer.nl/resources/Leefbaarheid in Nederland 2018.pdf

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C.C. Aggarwal & C.X. Zhai (eds.), *Mining Text Data*, 415-464. https://doi.org/10.1007/978-1-4614-3223-4_13

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015). Sentiment Analysis Techniques in Recent Works. In *2015 Science and Information Conference 2015 (SAI),* 288-291. https://doi.org/10.1109/SAI.2015.7237157

Manad, O., Menouer, T., & Darmon, P. (2019). Towards a Performant Multilingual Model Based on Ensemble Learning to Enhance Sentiment Analysis. *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA),* 1-8.

McCrea, R., & Walters, P. (2012). Impacts of Urban Consolidation on Urban Liveability: Comparing an Inner and Outer Suburb in Brisbane, Australia. *Housing Theory and Society,* 29(2), 190-206. https://doi.org/10.1080/14036096.2011.641261

McGreevy, M., Harris, P., Delaney-Crowe, T., Fisher, M., Sainsbury, P., Riley, E., & Baum, F. (2020). How well do Australian government urban planning policies respond to the social determinants of health and health equity? *Land Use Policy, 99*. https://doi.org/10.1016/j.landusepol.2020.105053

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PLoS ONE, 8(5). https://doi.org/10.1371/journal.pone.0064417

Municipality of Utrecht (2016, December 8). Ruimtelijke strategie 2016: Utrecht kiest voor gezonde groei. Retrieved from https://omgevingsvisie.utrecht.nl/fileadmin/uploads/documenten/zz-omgevingsvisie/2016-Ruimtelijke-strategie-Utrecht-kiest-voor-gezonde-groei.pdf

Namazi-Rad, M., Perez, P., Berryman, M., & Lamy, F. (2012). An experimental determination of perceived liveability in Sydney. In *ACSPRI Conferences, RC33 Eighth International Conference on Social Science Methodology,* 1-13.

Needham, C. J., Bradford, J. R., Bulpitt, A. J., Care, M. A., & Westhead, D. R. (2006). Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. *BMC Bioinformatics, 7,* 405.

Nenko, A., & Petrova, M. (2019). Comparing PPGIS and LBSN Data to Measure Emotional Perception of the City. In Communications in Computer and Information Science (Vol. 1038 CCIS, pp. 223–234). Springer. https://doi.org/10.1007/978-3-030-37858-5_18

Park, K. (2020). Social Capital and Residential Satisfaction in South Korea: A Comparative Study of Communities in Seoul, Yeoju and Gwacheon. *Environment and Urbanisation ASIA, 11*(1), 140-154. https://doi.org/10.1177/0975425320906279

Paul, A., & Sen, J. (2020). A critical review of liveability approaches and their dimensions. *Geoforum, 117*, 90-92.

Ročak, M., Hospers, G. J., & Reverda, N. (2016). Searching for social sustainability: The case of the shrinking city of Heerlen, the Netherlands. Sustainability (Switzerland), 8(4). https://doi.org/10.3390/su8040382

Rogers, A., Romanov, A., Rumshisky, A., Volkova, S., Gronas, M., Gribov, A. (2018). RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 201*8), 755–763.

Sabbadini, L. L., & Maggino, F. (2018). Quality of Life in Italian Official Surveys. Social Indicators Research, 135(3), 1043–1055. https://doi.org/10.1007/s11205-017-1766-2

Sarma, K. V., Spiegel, B. M. R., Reid, M. W., Chen, S., Merchant, R. M., Seltzer, E., & Arnold, C. W. (2019). Estimating the health-related quality of life of twitter users using semantic processing. *Studies in Health Technology and Informatics, 264*, 1065-1069. https://doi.org/10.3233/SHTI190388

Shafer, C. S., Koo Lee, B., & Turner, S. (2000). A tale of three greenway trails: user perceptions related to quality of life. *Landscape and Urban Planning, 49*(3-4), 163-178

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature communications, 9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

Staiano, J., & Guerini, M (2014). DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference, 2*, 427-433. https://doi.org/10.3115/v1/p14-2070

Statistics Netherlands (2021, March 31). Regionale kerncijfers Nederland. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?dl=52049.

Statistics Netherlands (2019a, September 10). Sterke groei in steden en randgemeenten verwacht. Retrieved from https://www.cbs.nl/nl-nl/nieuws/2019/37/sterke-groei-in-steden-en-randgemeenten-verwacht

Statistics Netherlands (2019b, December 17). Prognose: 19 miljoen inwoners in 2039. Retrieved from https://www.cbs.nl/nl-nl/nieuws/2019/51/prognose-19-miljoen-inwoners-in-2039

Suryowati, K., Bekti, R., & Faradila, A. (2018), A Comparison of Weights Matrices on Computation of Dengue Spatial Autocorrelation. *IOP Conference Series: Materials Science and Engineering, 355*(1), 1-9. https://doi.org/10.1088/1757-899X/335/1/012052

Tieskens, K. F., Van Zanten, B. T., Schulp, C. J. E., & Verburg, P. H. (2018). Aesthetic appreciation of the cultural landscape through social media: An analysis of revealed preference in the Dutch river landscape. *Landscape and Urban Planning, 177*, 128-137.

Tilaki, M. J. M., Bahauddin, A. A. A., & Marzbali, M. H. (2014). The Necessity of Increasing Livability for George Town Heritage Site: An Analytical Review. *Modern Applied Science, 8*(1), 123-133. http://dx.doi.org/10.5539/mas.v8n1p123

To, H., Agrawal, S., Kim, S. H., & Shahabi, C. (2017). On Identifying Disaster-Related Tweets: Matching-based or Learning-based? *Proceedings – 2017 IEEE 3rd International Conference on Multimedia Big Data,* 330-337. https://doi.org/10.1109/BigMM.2017.82

Treija, S., Bratuškins, U., Barvika, S., & Bondars, E. (2020). The liveability of historical cities: Current state and prospects for habitation. WIT Transactions on the Built Environment, 193, 15-26.

Tripaty, A., Agrawal, A, & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science, 57*, 821-829. https://doi.org/10.1016/j.procs.2015.07.523

Ubels., H., Bock, B., & Haartsen, T. (2019). An evolutionary perspective on experimental local governance arrangements with local governments and residents in Dutch rural areas of depopulation. *Environment and Planning C: Politics and Space, 37*(7), 1277-1295. https://www.doi.org/10.1177/2399654418820070

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterisation. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017,* 280-289.

Veenhoven, R. (1996). Hapy life-expectancy: A comprehensive measure of quality-of-life in nations. *Social Indicators Research, 39*, 1-58.

Venerandi, A. (2017). *A Quantitative Method to Study the Relationship between Urban Form and City Liveability Indexes* (Doctor of Engineering, University College London). Retrieved from https://discovery.ucl.ac.uk/id/eprint/1569319

Weichmann, T. & Bontje, M. (2015). Responding to Tough Times: Policy and Planning Strategies in Shrinking Cities. *European Planning Studies, 23*(1), 1-11. https://doi.org/10.1080/09654313.2013.820077

Wyatt, R. (2009). Heuristic Approaches to Urban Livability. *Malaysian Journal of Environmental Management, 10*(1), 43-65.

Zivanovic, S., Martinez, J., & Verplanke, J. (2020). Capturing and mapping quality of life using Twitter data. *GeoJournal, 85*(1), 237-255. https://doi.org/10.1007/s10708-018-9960-6

# 8. Appendices

## Appendix A: code gathering data

```
snscrape --jsonl twitter-search "since:2019-01-01 until:2019-12-31
geocode:52.169061048725325,5.480380931398393,225km" >Tweets19NL.json
```

The code used for preparing the data for Hydrator, pre-processing the Hydrator output, and the performance assessment can be found in the supplementary files or at https://github.com/MdeH1997/TwitterSA

## Appendix B: results performance assessment

| Lexicons | Classified (%) | Threshold | True positive | False positive | False negative | True negative | Accuracy | Precision | Recall | Specificity | $F_1$ | Youden's *J* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smedt | 40 | -0,30 | 63 | 26 | 0 | 5 | 0,723404 | 0,707865 | 1 | 0,16129 | 0,828947 | 0,16129 |
| | | -0,25 | 63 | 24 | 0 | 7 | 0,744681 | 0,724138 | 1 | 0,225806 | 0,84 | 0,225806 |
| | | -0,20 | 62 | 24 | 1 | 7 | 0,734043 | 0,72093 | 0,984127 | 0,225806 | 0,832215 | 0,209933 |
| | | -0,15 | 60 | 23 | 3 | 8 | 0,723404 | 0,722892 | 0,952381 | 0,258065 | 0,821918 | 0,210445 |
| | | -0,10 | 60 | 22 | 3 | 9 | 0,734043 | 0,731707 | 0,952381 | 0,290323 | 0,827586 | 0,242704 |
| | | -0,05 | 60 | 19 | 3 | 12 | **0,765957** | 0,759494 | 0,952381 | 0,387097 | **0,84507** | 0,339478 |
| | | 0,00 | 57 | 16 | 6 | 15 | **0,765957** | 0,780822 | 0,904762 | 0,483871 | 0,838235 | **0,388633** |
| | | 0,05 | 51 | 16 | 12 | 15 | 0,702128 | 0,761194 | 0,809524 | 0,483871 | 0,784615 | 0,293395 |
| | | 0,10 | 50 | 14 | 13 | 17 | 0,712766 | 0,78125 | 0,793651 | 0,548387 | 0,787402 | 0,342038 |
| | | 0,15 | 48 | 13 | 15 | 18 | 0,702128 | 0,786885 | 0,761905 | 0,580645 | 0,774194 | 0,34255 |
| | | 0,20 | 46 | 13 | 17 | 18 | 0,680851 | 0,779661 | 0,730159 | 0,580645 | 0,754098 | 0,310804 |
| | | 0,25 | 44 | 11 | 19 | 20 | 0,680851 | 0,8 | 0,698413 | 0,645161 | 0,745763 | 0,343574 |
| | | 0,30 | 41 | 11 | 22 | 20 | 0,648936 | 0,788462 | 0,650794 | 0,645161 | 0,713043 | 0,295955 |
| EmoWordNet | 97 | -0,30 | 153 | 74 | 2 | 0 | 0,668122 | 0,674009 | 0,987097 | 0 | 0,801047 | -0,012903 |
| | | -0,25 | 153 | 73 | 2 | 1 | **0,672489** | 0,676991 | 0,987097 | 0,013514 | **0,80315** | 0,00061 |
| | | -0,20 | 150 | 73 | 5 | 1 | 0,659389 | 0,672646 | 0,967742 | 0,013514 | 0,793651 | -0,018745 |
| | | -0,15 | 150 | 73 | 5 | 1 | 0,659389 | 0,672646 | 0,967742 | 0,013514 | 0,793651 | -0,018745 |
| | | -0,10 | 147 | 72 | 8 | 2 | 0,650655 | 0,671233 | 0,948387 | 0,027027 | 0,786096 | -0,024586 |
| | | -0,05 | 136 | 64 | 19 | 10 | 0,637555 | 0,68 | 0,877419 | 0,135135 | 0,766197 | 0,012554 |
| | | 0,00 | 112 | 43 | 43 | 31 | 0,624454 | 0,722581 | 0,722581 | 0,418919 | 0,722581 | 0,1415 |
| | | 0,05 | 71 | 23 | 84 | 51 | 0,532751 | 0,755319 | 0,458065 | 0,689189 | 0,570281 | 0,147254 |
| | | 0,10 | 51 | 8 | 104 | 66 | 0,510917 | 0,864407 | 0,329032 | 0,891892 | 0,476636 | **0,220924** |
| | | 0,15 | 25 | 3 | 130 | 71 | 0,419214 | 0,892857 | 0,16129 | 0,959459 | 0,273224 | 0,12075 |
| | | 0,20 | 10 | 3 | 145 | 71 | 0,353712 | 0,769231 | 0,064516 | 0,959459 | 0,119048 | 0,023976 |
| | | 0,25 | 4 | 1 | 151 | 73 | 0,336245 | 0,8 | 0,025806 | 0,986486 | 0,05 | 0,012293 |
| | | 0,30 | 2 | 1 | 153 | 73 | 0,327511 | 0,666667 | 0,012903 | 0,986486 | 0,025316 | -0,00061 |

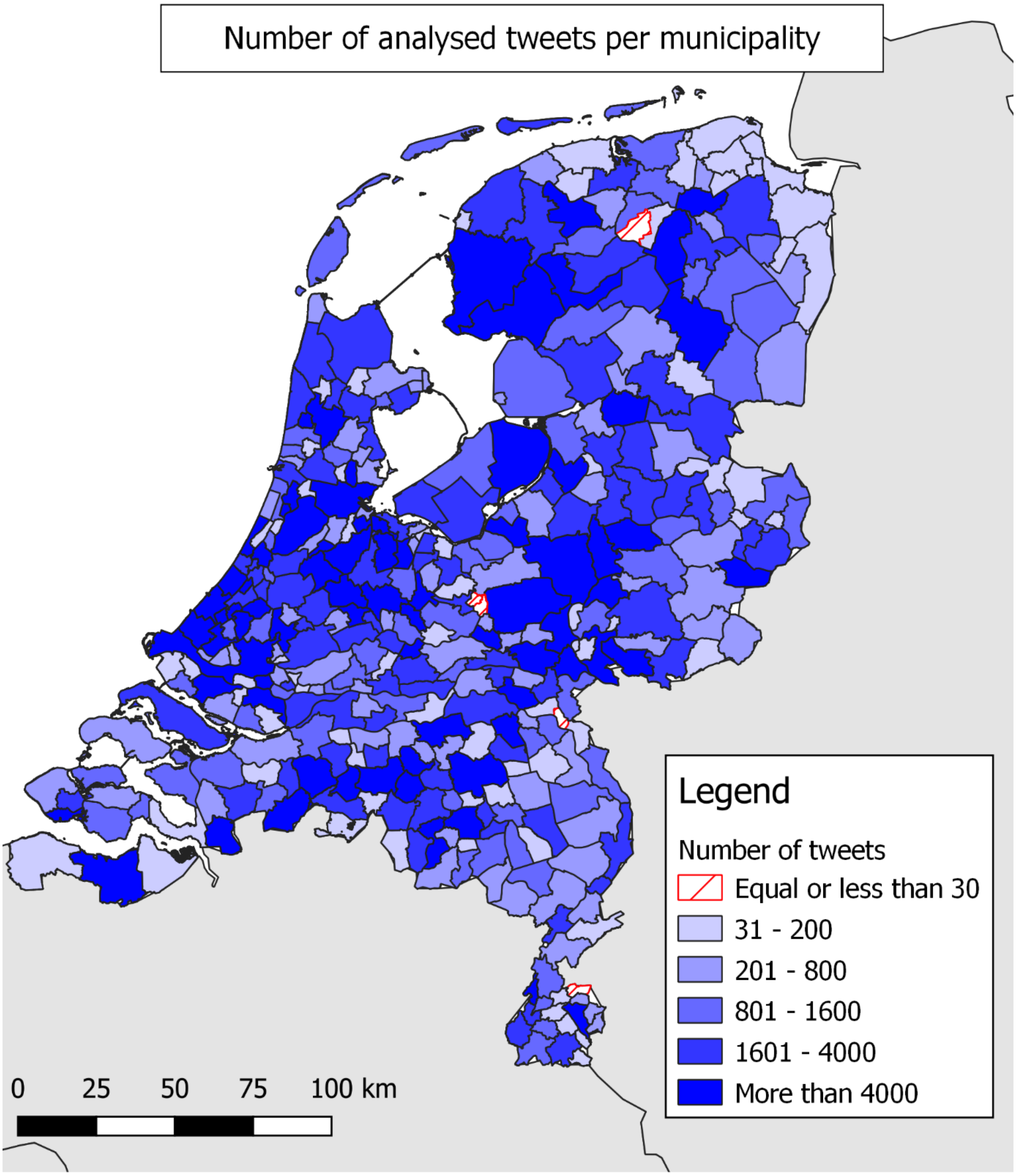
| Naïve Bayes’ | Classified (%) | Feature type | True positive | False positive | False negative | True negative | Accuracy | Precision | Recall | Specificity | $F_1$ | Youden’s *J* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dutch | 100 | plain text | 93 | 15 | 67 | 60 | 0,651064 | 0,861111 | 0,58125 | 0,8 | 0,69403 | 0,38125 |
| | | word, 1 | 93 | 9 | 67 | 66 | 0,676596 | 0,911765 | 0,58125 | 0,88 | 0,709924 | **0,46125** |
| | | word, 2 | 58 | 18 | 102 | 57 | 0,489362 | 0,763158 | 0,3625 | 0,76 | 0,491525 | 0,1225 |
| | | word, 3 | 48 | 10 | 112 | 65 | 0,480851 | 0,827586 | 0,3 | 0,866667 | 0,440367 | 0,166667 |
| | | word, 4 | 58 | 16 | 102 | 59 | 0,497872 | 0,783784 | 0,3625 | 0,786667 | 0,495726 | 0,149167 |
| | | word, 5 | 67 | 20 | 93 | 55 | 0,519149 | 0,770115 | 0,41875 | 0,733333 | 0,54251 | 0,152083 |
| | | character, 1 | 160 | 72 | 0 | 3 | 0,693617 | 0,689655 | 1 | 0,04 | **0,816327** | 0,04 |
| | | character, 2 | 135 | 43 | 25 | 32 | **0,710638** | 0,758427 | 0,84375 | 0,426667 | 0,798817 | 0,270417 |
| | | character, 3 | 120 | 39 | 40 | 36 | 0,66383 | 0,754717 | 0,75 | 0,48 | 0,752351 | 0,23 |
| | | character, 4 | 122 | 53 | 38 | 22 | 0,612766 | 0,697143 | 0,7625 | 0,293333 | 0,728358 | 0,055833 |
| | | character, 5 | 118 | 56 | 42 | 19 | 0,582979 | 0,678161 | 0,7375 | 0,253333 | 0,706587 | -0,009167 |
| English | 100 | plain text | 80 | 19 | 80 | 56 | 0,578723 | 0,808081 | 0,5 | 0,746667 | 0,617761 | 0,246667 |
| | | word, 1 | 78 | 14 | 82 | 61 | 0,591489 | 0,847826 | 0,4875 | 0,813333 | 0,619048 | **0,300833** |
| | | word, 2 | 78 | 21 | 82 | 54 | 0,561702 | 0,787879 | 0,4875 | 0,72 | 0,602317 | 0,2075 |
| | | word, 3 | 54 | 11 | 106 | 64 | 0,502128 | 0,830769 | 0,3375 | 0,853333 | 0,48 | 0,190833 |
| | | word, 4 | 75 | 20 | 85 | 55 | 0,553191 | 0,789474 | 0,46875 | 0,733333 | 0,588235 | 0,202083 |
| | | word, 5 | 160 | 75 | 0 | 0 | **0,680851** | 0,680851 | 1 | 0 | **0,810127** | 0 |
| | | character, 1 | 154 | 72 | 6 | 3 | 0,668085 | 0,681416 | 0,9625 | 0,04 | 0,797927 | 0,0025 |
| | | character, 2 | 116 | 39 | 44 | 36 | 0,646809 | 0,748387 | 0,725 | 0,48 | 0,736508 | 0,205 |
| | | character, 3 | 105 | 41 | 55 | 34 | 0,591489 | 0,719178 | 0,65625 | 0,453333 | 0,686275 | 0,109583 |
| | | character, 4 | 103 | 44 | 57 | 31 | 0,570213 | 0,70068 | 0,64375 | 0,413333 | 0,67101 | 0,057083 |
| | | character, 5 | 110 | 44 | 50 | 31 | 0,6 | 0,714286 | 0,6875 | 0,413333 | 0,700637 | 0,100833 |

| K-Nearest Neighbours | Classified (%) | Feature type | True positive | False positive | False negative | True negative | Accuracy | Precision | Recall | Specificity | $F_1$ | Youden's *J* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dutch | 100 | plain text | 109 | 34 | 51 | 41 | 0,638298 | 0,762238 | 0,68125 | 0,546667 | 0,719472 | 0,227917 |
| | | word, 1 | 100 | 20 | 60 | 55 | 0,659574 | 0,833333 | 0,625 | 0,733333 | 0,714286 | **0,358333** |
| | | word, 2 | 130 | 45 | 30 | 30 | 0,680851 | 0,742857 | 0,8125 | 0,4 | 0,776119 | 0,2125 |
| | | word, 3 | 158 | 68 | 2 | 7 | **0,702128** | 0,699115 | 0,9875 | 0,093333 | **0,818653** | 0,080833 |
| | | word, 4 | 160 | 75 | 0 | 0 | 0,680851 | 0,680851 | 1 | 0 | 0,810127 | 0 |
| | | word, 5 | 160 | 75 | 0 | 0 | 0,680851 | 0,680851 | 1 | 0 | 0,810127 | 0 |
| | | character, 1 | 69 | 26 | 91 | 49 | 0,502128 | 0,726316 | 0,43125 | 0,653333 | 0,541176 | 0,084583 |
| | | character, 2 | 73 | 10 | 87 | 65 | 0,587234 | 0,879518 | 0,45625 | 0,866667 | 0,600823 | 0,322917 |
| | | character, 3 | 77 | 11 | 83 | 64 | 0,6 | 0,875 | 0,48125 | 0,853333 | 0,620968 | 0,334583 |
| | | character, 4 | 104 | 22 | 56 | 53 | 0,668085 | 0,825397 | 0,65 | 0,706667 | 0,727273 | 0,356667 |
| | | character, 5 | 117 | 29 | 43 | 46 | 0,693617 | 0,80137 | 0,73125 | 0,613333 | 0,764706 | 0,344583 |
| English | 100 | plain text | 105 | 34 | 55 | 41 | 0,621277 | 0,755396 | 0,65625 | 0,546667 | 0,702341 | 0,202917 |
| | | word, 1 | 96 | 30 | 64 | 45 | 0,6 | 0,761905 | 0,6 | 0,6 | 0,671329 | 0,2 |
| | | word, 2 | 132 | 37 | 28 | 38 | **0,723404** | 0,781065 | 0,825 | 0,506667 | 0,802432 | **0,331667** |
| | | word, 3 | 152 | 65 | 8 | 10 | 0,689362 | 0,700461 | 0,95 | 0,133333 | 0,806366 | 0,083333 |
| | | word, 4 | 159 | 74 | 1 | 1 | 0,680851 | 0,682403 | 0,99375 | 0,013333 | 0,80916 | 0,007083 |
| | | word, 5 | 160 | 75 | 0 | 0 | 0,680851 | 0,680851 | 1 | 0 | **0,810127** | 0 |
| | | character, 1 | 88 | 28 | 72 | 47 | 0,574468 | 0,758621 | 0,55 | 0,626667 | 0,637681 | 0,176667 |
| | | character, 2 | 87 | 16 | 73 | 59 | 0,621277 | 0,84466 | 0,54375 | 0,786667 | 0,661597 | 0,330417 |
| | | character, 3 | 101 | 26 | 59 | 49 | 0,638298 | 0,795276 | 0,63125 | 0,653333 | 0,703833 | 0,284583 |
| | | character, 4 | 106 | 33 | 54 | 42 | 0,629787 | 0,76259 | 0,6625 | 0,56 | 0,70903 | 0,2225 |
| | | character, 5 | 112 | 37 | 48 | 38 | 0,638298 | 0,751678 | 0,7 | 0,506667 | 0,724919 | 0,206667 |

| Hybrid De Smedt | Classified (%) | Feature type | True positive | False positive | False negative | True negative | Accuracy | Precision | Recall | Specificity | $F_1$ | Youden's *J* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Naïve Bayes' | 99 | plain text | 103 | 26 | 55 | 48 | 0,650862 | 0,79845 | 0,651899 | 0,648649 | 0,71777 | 0,300547 |
| | | word, 1 | 107 | 21 | 51 | 53 | 0,689655 | 0,835938 | 0,677215 | 0,716216 | 0,748252 | **0,393431** |
| | | word, 2 | 98 | 26 | 60 | 48 | 0,62931 | 0,790323 | 0,620253 | 0,648649 | 0,695035 | 0,268902 |
| | | word, 3 | 93 | 25 | 65 | 49 | 0,612069 | 0,788136 | 0,588608 | 0,662162 | 0,673913 | 0,25077 |
| | | word, 4 | 99 | 31 | 59 | 43 | 0,612069 | 0,761538 | 0,626582 | 0,581081 | 0,6875 | 0,207663 |
| | | word, 5 | 108 | 35 | 50 | 39 | 0,633621 | 0,755245 | 0,683544 | 0,527027 | 0,717608 | 0,210571 |
| | | character, 1 | 152 | 56 | 6 | 18 | **0,732759** | 0,730769 | 0,962025 | 0,243243 | **0,830601** | 0,205269 |
| | | character, 2 | 134 | 40 | 24 | 34 | 0,724138 | 0,770115 | 0,848101 | 0,459459 | 0,807229 | 0,307561 |
| | | character, 3 | 123 | 38 | 35 | 36 | 0,685345 | 0,763975 | 0,778481 | 0,486486 | 0,77116 | 0,264967 |
| | | character, 4 | 125 | 40 | 33 | 34 | 0,685345 | 0,757576 | 0,791139 | 0,459459 | 0,773994 | 0,250599 |
| | | character, 5 | 121 | 42 | 37 | 32 | 0,659483 | 0,742331 | 0,765823 | 0,432432 | 0,753894 | 0,198255 |
| K-Nearest Neighbours | 99 | plain text | 121 | 39 | 37 | 35 | 0,672414 | 0,75625 | 0,765823 | 0,472973 | 0,761006 | 0,238796 |
| | | word, 1 | 118 | 31 | 40 | 43 | 0,693966 | 0,791946 | 0,746835 | 0,581081 | 0,76873 | 0,327917 |
| | | word, 2 | 135 | 43 | 23 | 31 | 0,715517 | 0,758427 | 0,85443 | 0,418919 | 0,803571 | 0,273349 |
| | | word, 3 | 150 | 53 | 8 | 21 | **0,737069** | 0,738916 | 0,949367 | 0,283784 | **0,831025** | 0,233151 |
| | | word, 4 | 152 | 59 | 6 | 15 | 0,719828 | 0,720379 | 0,962025 | 0,202703 | 0,823848 | 0,164728 |
| | | word, 5 | 152 | 59 | 6 | 15 | 0,719828 | 0,720379 | 0,962025 | 0,202703 | 0,823848 | 0,164728 |
| | | character, 1 | 98 | 34 | 60 | 40 | 0,594828 | 0,742424 | 0,620253 | 0,540541 | 0,675862 | 0,160794 |
| | | character, 2 | 107 | 23 | 51 | 51 | 0,681034 | 0,823077 | 0,677215 | 0,689189 | 0,743056 | **0,366404** |
| | | character, 3 | 106 | 25 | 52 | 49 | 0,668103 | 0,80916 | 0,670886 | 0,662162 | 0,733564 | 0,333048 |
| | | character, 4 | 113 | 32 | 45 | 42 | 0,668103 | 0,77931 | 0,71519 | 0,567568 | 0,745875 | 0,282757 |
| | | character, 5 | 122 | 38 | 36 | 36 | 0,681034 | 0,7625 | 0,772152 | 0,486486 | 0,767296 | 0,258638 |

| Hybrid EmoWordNet | Classified (%) | Feature type | True positive | False positive | False negative | True negative | Accuracy | Precision | Recall | Specificity | $F_1$ | Youden's *J* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Naïve Bayes' | 100 | plain text | 53 | 8 | 107 | 67 | 0,510638 | 0,868852 | 0,33125 | 0,893333 | 0,479638 | 0,224583 |
| | | word, 1 | 54 | 8 | 106 | 67 | 0,514894 | 0,870968 | 0,3375 | 0,893333 | 0,486486 | 0,230833 |
| | | word, 2 | 54 | 9 | 106 | 66 | 0,510638 | 0,857143 | 0,3375 | 0,88 | 0,484305 | 0,2175 |
| | | word, 3 | 55 | 9 | 105 | 66 | 0,514894 | 0,859375 | 0,34375 | 0,88 | 0,491071 | 0,22375 |
| | | word, 4 | 55 | 9 | 105 | 66 | 0,514894 | 0,859375 | 0,34375 | 0,88 | 0,491071 | 0,22375 |
| | | word, 5 | 55 | 9 | 105 | 66 | 0,514894 | 0,859375 | 0,34375 | 0,88 | 0,491071 | 0,22375 |
| | | character, 1 | 55 | 8 | 105 | 67 | **0,519149** | 0,873016 | 0,34375 | 0,893333 | **0,493274** | **0,237083** |
| | | character, 2 | 54 | 9 | 106 | 66 | 0,510638 | 0,857143 | 0,3375 | 0,88 | 0,484305 | 0,2175 |
| | | character, 3 | 55 | 8 | 105 | 67 | **0,519149** | 0,873016 | 0,34375 | 0,893333 | **0,493274** | **0,237083** |
| | | character, 4 | 54 | 9 | 106 | 66 | 0,510638 | 0,857143 | 0,3375 | 0,88 | 0,484305 | 0,2175 |
| | | character, 5 | 55 | 9 | 105 | 66 | 0,514894 | 0,859375 | 0,34375 | 0,88 | 0,491071 | 0,22375 |
| K-Nearest Neighbours | 100 | plain text | 53 | 8 | 107 | 67 | 0,510638 | 0,868852 | 0,33125 | 0,893333 | 0,479638 | 0,224583 |
| | | word, 1 | 54 | 8 | 106 | 67 | **0,514894** | 0,870968 | 0,3375 | 0,893333 | 0,486486 | **0,230833** |
| | | word, 2 | 55 | 9 | 105 | 66 | **0,514894** | 0,859375 | 0,34375 | 0,88 | **0,491071** | 0,22375 |
| | | word, 3 | 55 | 9 | 105 | 66 | **0,514894** | 0,859375 | 0,34375 | 0,88 | **0,491071** | 0,22375 |
| | | word, 4 | 55 | 9 | 105 | 66 | **0,514894** | 0,859375 | 0,34375 | 0,88 | **0,491071** | 0,22375 |
| | | word, 5 | 55 | 9 | 105 | 66 | **0,514894** | 0,859375 | 0,34375 | 0,88 | **0,491071** | 0,22375 |
| | | character, 1 | 55 | 9 | 105 | 66 | **0,514894** | 0,859375 | 0,34375 | 0,88 | **0,491071** | 0,22375 |
| | | character, 2 | 53 | 8 | 107 | 67 | 0,510638 | 0,868852 | 0,33125 | 0,893333 | 0,479638 | 0,224583 |
| | | character, 3 | 54 | 8 | 106 | 67 | **0,514894** | 0,870968 | 0,3375 | 0,893333 | 0,486486 | **0,230833** |
| | | character, 4 | 54 | 9 | 106 | 66 | 0,510638 | 0,857143 | 0,3375 | 0,88 | 0,484305 | 0,2175 |
| | | character, 5 | 54 | 9 | 106 | 66 | 0,510638 | 0,857143 | 0,3375 | 0,88 | 0,484305 | 0,2175 |

## Appendix C: Number of tweets per municipality



Municipalities with less than 30 classified tweets:

| Municipality | Number of tweets |
|---|---|
| Onderbanken | 13 |
| Renswoude | 18 |
| Marum | 26 |
| Mook en Middelaar | 26 |
| Scherpenzeel | 29 |

## Appendix D: Outliers

Municipalities with a sentiment score of more than 3 standard deviations from the mean are treated as outliers. Table H.1 shows the municipalities with their corresponding scores and the number of classified tweets this score is composed of.

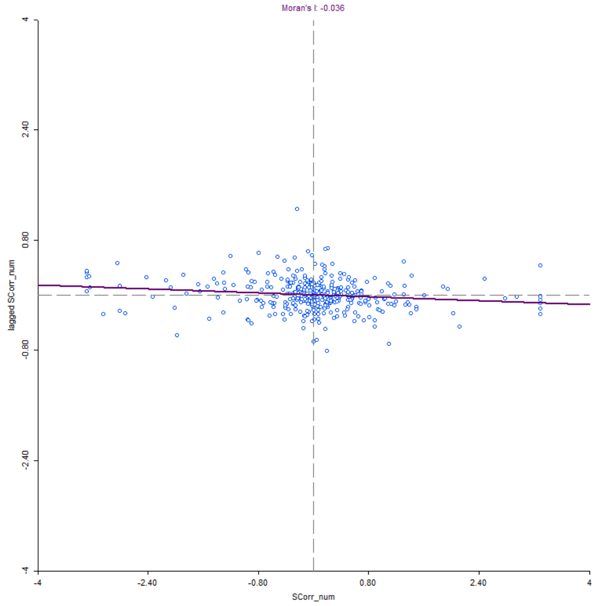
| Municipality | Sentiment score | Number of tweets |
|---|---:|---:|
| Reusel-De Mierden | -0,578 | 33 |
| Onderbanken | -0,538 | 13 |
| Doesburg | -0,485 | 66 |
| Aalten | -0,448 | 87 |
| Geertruidenberg | 0,045 | 44 |
| Rozendaal | 0,050 | 40 |
| Hulst | 0,055 | 91 |
| Baarle-Nassau | 0,064 | 47 |
| Nuth | 0,091 | 44 |
| Renswoude | 0,111 | 18 |

## Appendix E: Spatial analysis sentiment scores

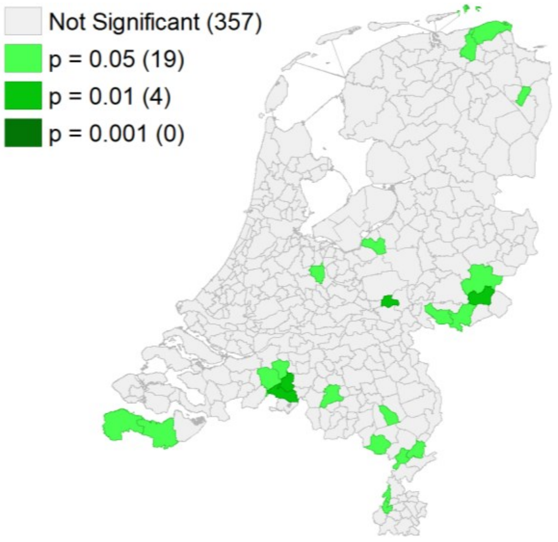
The expected *I* is -0,0026. Based on 999 permutations, the 2nd order Queen's contiguity shows significant spatial autocorrelation.

| Neighbourhood matrix | | | | Global Moran's I | | |
|---|---|---|---|---|---|---|
| | | Number of neighbours | | | | |
| Type | Specification | Min – max | Mean | Moran's *I* | Pseudo p-value | Z-value |
| Queen's contiguity | 1st order | 1 – 18 | 5,30 | -0,016 | 0,338 | -0,390 |
| | 2nd order | 2 – 38 | 16,93 | -0,036 | 0,025 | -1,809 |
| | 3rd order | 3 – 62 | 34,26 | -0,007 | 0,364 | -0,364 |

Global Moran's *I* scatterplot of sentiment scores and lagged sentiment scores



Local Moran's *I* Significance of sentiment score clusters.


Not Significant (357)
p = 0.05 (19)
p = 0.01 (4)
p = 0.001 (0)