Analysis of tweets in relation to climate change based on the geographical origin of the sender

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Michael Kirchknopf, BSc
Matrikelnummer 01126331

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Mag. Dr. Dieter Merkl

Wien, 1. März 2023

[redacted]

Michael Kirchknopf	Dieter Merkl

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Informatics

Analysis of tweets in relation to climate change based on the geographical origin of the sender

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Michael Kirchknopf, BSc
Registration Number 01126331

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Mag. Dr. Dieter Merkl

Vienna, 1st March, 2023

[redacted]

Michael Kirchknopf

Dieter Merkl

Erklärung zur Verfassung der Arbeit

Michael Kirchknopf, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. März 2023

Michael Kirchknopf

Danksagung

Zu Beginn der vorliegenden Diplomarbeit möchte ich mich bei all jenen bedanken, die mich auf unterschiedlichste Art und Weise während meines Studiums begleitet und bei der Verfassung dieser Diplomarbeit unterstützt haben.

Besonderer Dank gilt dabei meinem Betreuer, Ao.Univ.Prof. Mag. Dr. Dieter Merkl, der mich bei dieser Arbeit stets unterstützt hat. Sei es bei der Unterstützung in der Themenfindung oder bei allfälligen Fragen, seine Meinung und sein Wissen haben einen großen Beitrag zu dieser Diplomarbeit beigetragen. Aber auch die Unterhaltungen abseits der Diplomarbeit haben für eine sehr angenehme Atmosphäre gesorgt. All dies hat mir die Verfassung dieser Arbeit wesentlich erleichtert.

Zusätzlich gilt ein großer Dank meiner Familie, die mich während meiner Studienjahre in den unterschiedlichsten Aspekten unterstützt und mir dadurch dieses Studium ermöglicht haben. Ebenfalls möchte ich mich auch bei meiner Freundin und meinen Freunden bedanken, ohne deren Antrieb und Motivation mir diese Diplomarbeit sicherlich deutlich schwerer gefallen wäre.

Kurzfassung

Der Klimawandel ist ein fortschreitendes globales Problem, weshalb sich viele Menschen darüber eine Meinung bilden und diese in sozialen Netzwerken teilen sowie diskutieren. Twitter ist eine der am häufigsten genutzten Microblogging Plattformen und bietet daher ein breitgestreutes Meinungsbild. Da der Klimawandel unterschiedliche Auswirkung auf Menschen in unterschiedlichen Regionen hat, ist das Ziel dieser Arbeit die Unterschiede von Tweets in geografisch unterschiedlich gelegenen Städten zu untersuchen und zu vergleichen. Dabei wurden Twitter Daten aus den Jahren 2020 und 2021 von den europäischen Hauptstädten hinsichtlich der Stimmungslage und polarisierenden Wörtern miteinander verglichen. Das Hauptaugenmerk lag dabei herauszufinden, ob es Unterschiede in den Tweets von Usern in Städten an der Küste zu Usern in Städten im Inland gibt. Die Analyse und Auswertung wurde anhand von CRISP-DM, einem technologie- und branchenunabhängigen Prozessmodell für Data Mining, durchgeführt. Bei dieser Arbeit wurde ein Ansatz gewählt, der es auf Grundlage des angegebenen Standorts des Nutzerprofils ermöglicht Tweets europäischen Hauptstädten zuzuordnen. Dadurch konnten wesentlich mehr Tweets für die Analyse herangezogen werden, als bei Analysen von Tweets auf Basis der GPS-Daten. Mit dem gesammelten Datensatz wurde ein Vergleich zwischen europäischen Hauptstädten an der Küste und im Inland im Bereich des Klimawandels durchgeführt. In diesem Vergleich konnten bei polarisierenden Wörtern Unterschiede zwischen den Gruppen von Städten an der Küste und Städten im Inland festgestellt werden. Bei der Stimmungslage konnten keine Unterschiede festgestellt werden. Überraschenderweise war die allgemeine Stimmung leicht positiv, während man bei diesem Thema eigentlich eine negative Stimmung erwarten könnte.

Abstract

Climate change is an increasing global problem, which is why many people are forming opinions about it by sharing and discussing them on social networks. Twitter is one of the most widely used microblogging platforms and therefore offers a broad range of opinions. Since climate change has different impacts on people in different regions, the aim of this thesis is to investigate and compare the differences of tweets in geographically different cities. For this purpose, Twitter data from 2020 and 2021 from European capitals were compared in terms of sentiment and polarizing words. The main focus was to find out if there are differences in tweets from users in coastal cities compared to users in inland cities. The analysis and evaluation was performed using CRISP-DM, a technology and industry independent process model for data mining. In this work, an approach was chosen that enabled tweets to be assigned to European capitals based on the specified location in the user profile. This allowed many more tweets to be used for analysis than if only tweets with GPS data had been used. With the gathered dataset a comparison was made between coastal and inland European capitals in the field of climate change. In this comparison, differences were found between the groups of coastal and inland cities for polarizing words, but not for sentiment. Surprisingly, the overall sentiment was slightly positive, whereas one might actually expect a negative sentiment on this topic.

Contents

CHAPTER 1

Introduction

1.1 Problem Statement and Aim of the Work

Climate change is a very defining topic and concerns people all over the world. Global warming continues to progress [Bro20, MIB21] and the frequency of alarming reports is increasing. However, the impact of global warming has different effects on different cities. Especially on cities near the coast, climate change has major consequences (e.g., rising sea levels due to polar melting) [Kum21]. But other factors such as glacier melting, droughts, floods can also affect life in cities. The large number of postings on social media platforms such as Twitter show that these topics concern humankind.

Twitter is a platform where one can get a good picture of the mood and polarizing hashtags and words on various topics. Especially when it comes to climate change, there is a large number of different hashtags under which opinions on climate change are expressed. Two hashtags in particular dominate discussions on this topic on Twitter: #climatechange and #globalwarming [CRM+15, SFW+20]. The hashtag #climatechange is used more in scientific contexts and the hashtag #globalworming in political topics and topics related to environmental disasters [SFW+20] and therefore has a more negative effect [CRM+15].

The challenge in analyzing tweets is to cleanly prepare the data so that it is usable for processing and analysis. Not all tweets with common hashtags are useful for analysis or may be in different languages [DKL19]. Information about the origin of a user can also be entered textually, resulting in different versions of the same city. For example, Vienna can be in common usage 'Vienna', 'Wien' or 'Vienna, Austria' among others. However, users also have the option of adding GPS data to their tweets, but this is only done within a range of 0,42% to 3,17% of tweets [KS14, LRK16, STC21], tending towards the lower limit. With geo-tagging one has the exact location of the user when they post their texts, but this does not always have to relate to their origin (e.g. vacation or business trips). When analyzing only geo-tagged tweets, a lot of interesting information is lost from tweets that are not geo-tagged, since the number of geo-tagged information is small. Therefore, localizing using other information, such as the user profile location, is a good alternative [LRK16].

The target of this work is to gather tweets in the context of climate change and information about cities to join them based on the user profile location of Twitter users to then identify the current mood and polarizing words in the context of climate change on a city level. A comparison within coastal cities and within inland cities aims to show whether there are similarities in these groups and whether these groups differ from each other. This results in the following research questions:

1. Which terms in the context of climate change polarize on Twitter in European cities on the coast and in European inland cities and do these groups differ?
2. Which three European cities have the most negative mood about climate change based on using the sentiment of tweets compared to other European cities and where are these cities located at?
3. Are there similarities in the context of climate change within European cities on the coast and within European inland cities based on using the sentiment of tweets and do these groups differ?

The goal of the project is to collect data from Twitter from a certain period of time with certain hashtags related to climate change in English language. In order to find out which time period and which hashtags should be used, an initial analysis and comparison with results from the literature is needed, as there are a lot of different hashtags in the field of climate change, with some being more and some less clearly attributable to climate change [HZ15]. The goal is then to find a way to assign the tweets of the selected dataset to European cities based on the information contained in the user profile and analyse it with text mining techniques to capture different sentiments of different cities and what is relevant in each of these cities in relation to climate change. The allocation based on the user profile location should allow a significantly larger number of tweets to be considered for a period of time than when using GPS data. European cities only make up a certain proportion of all tweets, which means that a sufficient amount of data is essential in order to provide meaningful answers to the research questions.

In the analysis of the data the objective is to find out if there are noticeable differences in the tweeting behavior in European cities between coastal cities and inland cities. The circumstances of these two groups are different, as coastal cities are affected by different issues than inland cities. Various sentiment analyses are to be tested in order to find the model that fits best. As output the sentiment will be about equalized comparing tweets near the coast and inland, because negative events of various kinds can occur anywhere. Since tweets usually follow current events [DKL19], it is expectable that there are different issues that cities are concerned about and thus similarities occur among similarly situated cities. In the end, the aim is to provide an overview of the mood and polarizing words related to climate change in Europe to determine what is keeping users concerned on Twitter in the respective cities.

1.2 Research Method

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a process model which provides a framework for data mining projects [WH00]. It is a technology and industry independent model [WH00, SKG21, MPCOF$^+$21]. CRSIP-DM was developed by several companies in the 1990s [CCK$^+$00, KM06] and officially proposed by Wirth & Hipp in 2000 [WH00, KM06, MSMFB09]. Until then, there has been no standard framework for data mining projects [WH00, MSMFB09]. In the 1990s several attempts were made in the field of Knowledge Discovery and Data Mining (KDDM) to define methodologies and therefore CRSIP-DM builds on these approaches [WH00]. CRSIP-DM has established itself over the years as KDDM and is the most widely used cycle framework in this area [KM06, Sal21, MPCOF$^+$21]. It has become the de-facto standard in data mining projects [SKG21, MSMFB09, MPCOF$^+$21] and is still the most widely used framework today [Sal21, MPCOF$^+$21]. However, there are also several models and methodologies which are specializations, extensions or enrichments of CRSIP-DM [MPCOF$^+$21, Sal21].

CRSIP-DM consists of several phases, which themselves are divided into tasks and deliverables. It has the goal to make data mining projects cheaper, more reliable and more repeatable, so that these projects can be carried out faster and more clearly. CRISP-DM is a guide through the project and helps to structure and subsequently document the process [WH00, CCK$^+$00]. The phases within CRISP-DM are represented as a life cycle, as shown in figure 1.1.

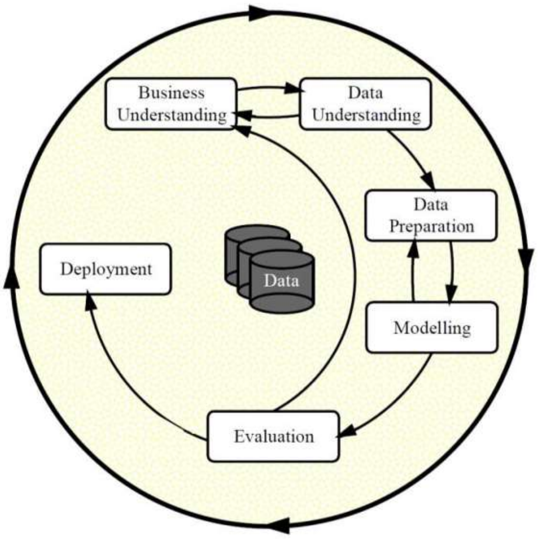
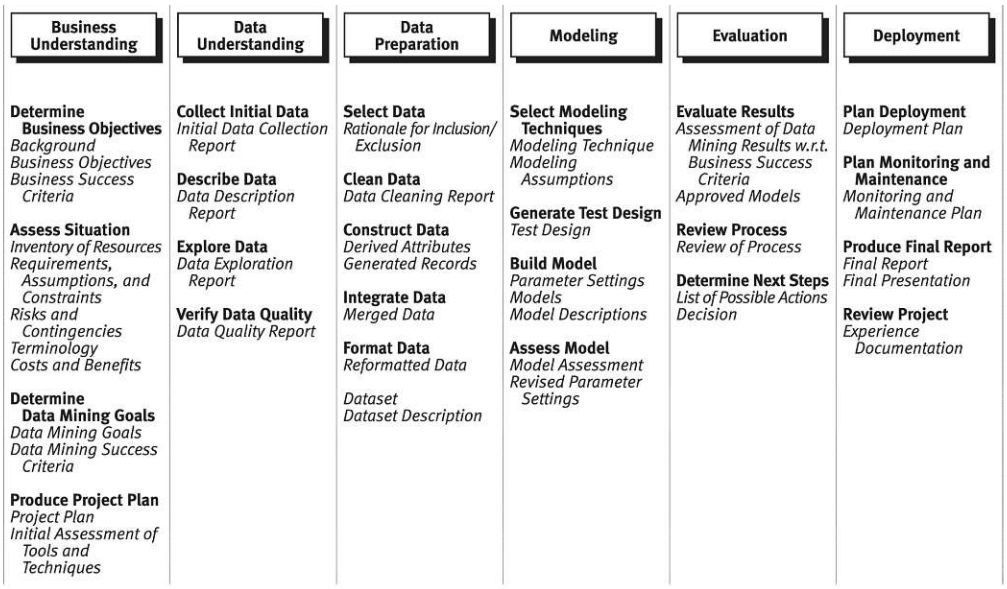


Figure 1.1: Phases of CRISP-DM[WH00]

The life cycle of CRISP-DM consists of six phases, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [WH00, CCK+00]. These phases are processed sequentially, whereby moving back is possible [CCK+00]. This depends on the results of the individual phases, which can make it necessary to make adjustments in previous phases [CCK+00]. The arrows indicate the main dependencies between the phases [WH00]. The outer cycle symbolizes the field of data mining, since data mining projects can be repeated several times [WH00].

Within these phases, CRSIP-DM consists of different tasks that are performed during the process of data mining projects and these tasks lead to several outputs of the respective tasks[Sal21]. In figure 1.2 the phases are broken down into the tasks and outputs.



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 1.2: Overview of the CRISP-DM tasks and their outputs[WH00, CCK+00]

In the following the six phases and its tasks and outputs are briefly summarized [WH00, CCK+00, Sal21].

1. **Business Understanding:** This phase is about understanding the objectives, requirements and challenges of the project from a business perspective. To do this, the initial situation must be understood, including what resources are available or what risks there are. This results in the business objectives and the project goals. To structure the work, a project plan is created by defining technologies and tools for each phase. From this phase, business requirements, assumptions, success criteria and risks result as output.

2. **Data Understanding:** In this phase, the first data is collected and subsequently analyzed to get familiar with the data. There is a close link between this and the previous phase, since information from the data is already required for the

deliverables in the Business Understanding phase. Furthermore, in the Data Understanding phase the data is described by means of field properties and data types. In order to successfully process the data, data quality problems are identified and analyzed. The output of this phase is an initial dataset, a description of the data and the characteristics and problems found.

3. **Data Preparation:** Data preparation is about creating a final dataset for the remaining phases. Usually, this phase is the most time-consuming of all phases, as many tasks have to be executed several times to get a proper result. The tasks in this phase can be, among others, cleaning the data, formatting, adding new attributes and integrating data from different sources. The output is a final dataset that eliminates the previously found data quality problems and is suitable for the Modeling phase.

4. **Modeling:** In the Modeling phase, different methods and models are selected and applied to find the appropriate methods and models to answer the data mining goals. These methods and models are used to experiment with the parameters and to compare them with each other in order to find a suitable model for the evaluation. Here exists a close link with the Data Preparation phase, since one might encounter new data problems in the Modeling phase and these have to be corrected in the Data Preparation phase. The result of this phase is the selection of a model and its parameter settings.

5. **Evaluation:** In this phase, the model is evaluated with regard to the business objectives to determine whether it meets the requirements and can answer them. For this purpose, the results are evaluated and the process is reviewed. The output of this phase is a summary of the results and whether these results have achieved the data mining goals. Here the decision has to be made whether to continue with these results into the Deployment phase or whether to go back to the previous phase.

6. **Deployment:** The Deployment phase is about presenting or using the results. This can be a report or an implementation of a repeatable data mining process.

1.3 Methodological Approach

This work is divided into the six phases of the CRISP-DM life cycle. In the phases, the data is understood, processed and evaluated step by step. In the thesis the six phases of CRISP-DM are followed as the following:

1. **Business Understanding:** In order to understand business and project objectives and its challenges, it is important to first get an overview of the possible characteristics of data. For this purpose, data from 01.01.2020 until 31.12.2021 in English are subtracted, which contain the hashtags #climatechange, #globalwarming, #climate

or combinations of these. This leads to a dataset with which initial analyses can be made. In order to understand the behavior of Twitter users, the first step is to analyze the occurrences of hashtags with the help of literature examples to make a more precise selection of hashtags. In the following, analyses of the usage of sharing locations in tweets and usage of user profile location is done, so that a comparison of the amount of data can be made. Through these analyses, the selection of parameters for further datasets is made and the data mining goals to answer the research questions are defined.

2. **Data Understanding:** After the first insights into Twitter have been delivered in the Business Understanding phase, it is then a matter of deepening these insights and looking at concrete cases in the dataset. To do this, a script is needed that extracts the data with the predefined parameters. In order to better understand the results, the data is described in further detail. Then, data is loaded into a database and deeper analysis about hashtags, user profile location, users and texts is done, which brings new insights from the data and identifies data quality problems in it. These problems must be resolved in the next phase (Data Preparation). In the Data Understanding phase not only tweets are gathered and analyzed, but also a city dataset with European capitals containing information (e.g. distance to the sea) to help answer the research questions.

3. **Data Preparation:** This phase is about preparing the data so that it can be analyzed in the next phases. In the Data Understanding phase, two datasets are collected, one with Twitter data (e.g. texts, hashtags, users) and one with European capitals (e.g. city name, country, distance to sea). Both datasets are separate tables that have to be joined together. For this, a suitable way must be found to link the cities with the tweets using the user profile location field. Furthermore, several text preparation steps have to be performed, which includes among others the removal of hashtags, hyperlinks and punctuations, so that disturbing information for data analysis is removed. Additionally, tweets from e.g. bots are filtered out to remove tweets from users with high similarity.

4. **Modeling:** At the beginning of this phase, a final dataset prepared and pre-analyzed in the previous phases is available. This dataset is processed in this phase using methods and models. Depending on the research question, appropriate methods and models are selected to answer the research questions then in the next phase. Thereby the two topics polarizing words and sentiment analysis are covered. For the topic of polarizing words word clouds and relative comparison of texts are discussed. In the case of sentiment analysis, various sentiment analyses are considered and discussed in order to make a selection.

5. **Evaluation:** In this phase, the results of the two topics (polarizing words and sentiment analysis) are evaluated per European capital in order to make comparisons with each other. For the sentiment analysis a ranking is created to show the cities with the most negative sentiment. In addition, the ranking can be used to see

whether similarities can be found within European cities on the coast and within European inland cities. Therefore the capitals were first classified as groups (inland cities and coastal cities) and then compared. In order to find out which terms polarize in European cities, word clouds are created per city as visualization. In making comparisons of the data between the European capitals, relative comparisons of the data are made because each city has different amounts of data in the dataset. This way they can be compared with each other to find out if they show more similarities within the same group than the other group.

6. **Deployment:** In this phase the results of the sentiment analysis and the analysis of polarizing words are visually presented in a map of Europe where the levels of similarity and sentiment are considered.

1.4 State of the Art

Twitter is a worldwide used social media platform in the field of microblogging. Twitter allows users to send posts (so-called tweets) with a length of up to 280 characters on any topic, such as climate change. With the help of hashtags, words or words chained together are recognized as topics and discussed on Twitter (e.g. #climatechange or #globalwarming) [SFW+20]. Since tweets express the thoughts and reflections of users on certain topics, especially through real-world events, it sparks the interest of social scientists to take advantage of this data [DKL19]. Especially the large amount of data on microblogging platforms like Twitter on all kinds of topics makes such platforms valuable for academic researches [KS14].

Studies in the field of climate change (natural phenomena such as weather or climate) use Twitter as a source of information because the reaction to occurrences of events is posted there and the large amount of tweets can be analysed using data mining techniques [KS14]. Different APIs and libraries in various programming languages allow tweets to be mined for a very large amount of data for relatively little effort. Thus, information such as tweet texts, hashtags, user profile descriptions, geo-tagging, likes and much more can be extracted, offering a wide range of data mining possibilities.

In the field of climate change on Twitter, there are different approaches to analyse tweets. For example, analyses have already been done to compare tweets with different hashtags [CRM+15, SFW+20], but also various approaches in the area of geo-location assignment, based on geo-tagging [STC21] and approaches to localization based on other existing data, such as textual contents of users [DKL19, LRK16]. The results show that there is still much room for further possibilities in this field. In the sentiment analysis area, Biraj Dahal et al [DKL19] use global data to show the temporal and geospatial perspectives in climate change discussions on Twitter and the challenges of Twitter datasets.

Since the analysis of tweets mostly involves text analysis, text mining techniques such as in the field of Natural Language Processing (NLP) are used. NLP describes techniques and methods for machine processing of natural language, such as sentiment analysis

[HG14, MSM22]. Due to free texts from users, there are challenges with sentiment analysis on Twitter, as texts can contain spelling errors and grammar mistakes. However, a lot of effort is put into research to keep improving sentiment analysis on Twitter [CP21]. In addition word clouds are a good way of visualizing information about the frequencies of words in texts [HLLE14]. They help to quickly find out from texts which words are of importance in a context and can be used to draw conclusions and support further analysis.

CHAPTER 2

Business Understanding

In order to understand information in a context, it is important to first get an overview of the possible characteristics of data. Especially on Twitter there are a lot of possibilities to share information and thoughts, since almost all information is text-based. Therefore, in order to make a clean analysis, one should first understand the possibilities of Twitter users to select data sets for deeper analysis.

To have a meaningful dataset to analyse climate change data later on, various initial analyses are done using a sample of data. The conditions for the data sample are made based on initial approximations of a possible data set. So the following conditions are selected to serve as the basis for the analysis and findings in this phase:

- tweets from 01.01.2020 until 31.12.2021
- tweets with the hashtags #climatechange, #globalwarming, #climate or combinations of these
- tweets in English language

This results in 2.510.748 tweets. With those tweets, analysis and insights will determine the data set for the analysis to answer the research questions and thus the data mining goals.

2.1 Hashtags in relation to Climate Change in Tweets

There are many hashtags on Twitter that can be associated with climate change. A first choice fell on the hashtags #climatechange, #globalwarming and #climate. In particular the hashtags #climatechange and #globalwarming dominate discussions on this topic on Twitter[CRM+15, SFW+20]. In contrast, the hashtag #climate is more general, can be ambiguous and is less significant for climate change discussions on Twitter [HZ15]. However, this hashtag can be good for identifying other hashtags related to climate change and therefore it was selected to gather more insights. It has been shown that more complex hashtags are more significant in relation to climate change than less complex ones [HZ15]. This means that the combination of words like climate and change to #climatechange or global and warming to #globalwarming can be assigned more significantly to climate change than individual words.

When evaluating the top 10 hashtags based on occurrences in this data sample, one can see that the three preselected hashtags are of course in the top 3, see table 2.1. Of these three hashtags, #climatechange is clearly the most used hashtag during this period and is used about nine times as often as #globalwarming. The hashtag #climatechange is used more in scientific contexts and the hashtag #globalworming in political topics and topics related to environmental disasters [SFW+20] and therefore has a more negative effect [CRM+15].

Hashtag	**Occurrences**
#climatechange	1.813.856
#climate	676.866
#globalwarming	219.239
#climatecrisis	204.299
#climateaction	195.778
#environment	126.506
#sustainability	113.815
#climateemergency	110.801
#cop26	97.144
#covid19	65.734

Table 2.1: Top 10 hashtags in the dataset

The goal with these three initial hashtags was to identify new hashtags, so the other seven in this table are more interesting in this analysis. Not every hashtag found in this analysis is related to climate change. A good example is #covid19. In these two selected years, Covid19 was of course a dominant topic alongside climate change, but this hashtag doesn't necessarily have anything to do with climate change, although it was used with hashtags related to climate change. This hashtag was linked with climate change topics in these tweets, but one cannot conclude from this that all tweets with #covid19 have something to do with climate change. The same applies to #environment and #sustainability. In tweets with other hashtags related to climate, these make sense, but if they stand alone in a tweet, one could not be able to conclude for sure that they are related to climate change.

On the other hand, there are also hashtags in table 2.1 of the top 10 hashtags that, according to the results of Marina L. Gavrilova et al [HZ15], are more significantly associated with climate change. The hashtags to which this applies are #climatecrisis, #climateaction and #climateemergency and are descriptive and clearly attributable to climate change. One additional hashtag appears in table 2.1 of the top 10 hashtags, namely #cop26. COP26 was the 26th Conference of the Parties (COP) to the United Nations Framework Convention on Climate Change (UNFCCC) and took place in Glasgow from 31st of October to 13th of November 2021 [LM22]. So #cop26 also relates to climate change. The conference before (COP25) was in 2019 and therefore is not as present in the data as COP26.

In reviewing more hashtags, other interesting hashtags have come up, as shown in table 2.2. The hashtags #climatechangeisreal and #climatehoax could hardly represent a more different opinion. While #climatechangeisreal represents climate change believers, #climatehoax represents disbelievers [TBCS20]. The others (#climatejustice, #climatebrawl and #climatestrike) are more general hashtags and cover areas which not fully covered by the top 10 hashtags from table 2.1.

Hashtag	**Occurrences**
#climatechangeisreal	23.334
#climatejustice	19.338
#climatebrawl	19.290
#climatestrike	13.367
#climatehoax	6.951

Table 2.2: Other hashtags in the dataset

2.2 Usage of Sharing Location in Tweets

Twitter allows users to post tweets with exact location. However, this option is disabled by default and must be enabled manually by users. This is one of the reasons why this feature is very rarely used. With this feature enabled, location data can be shared with tweets. This option can be turned on or off at any time.

There have already been different evaluations of various scientific works that have come to a range of 0,42% to 3,17% with shared GPS data in their analyzed datasets [KS14, LRK16, STC21]. Thus, by analysing only these tweets a lot of interesting information is lost because of a huge amount of users are not sharing their location.

In the dataset used in this phase for analysis, there are 2.510.748 tweets and of those tweets 57.915 share a location. So 2,31% tweets of the dataset share their location. Of those 57.915 tweets, 2.380 tweets can be found linked to European capitals. For example in Vienna there are 78 tweets within this dataset. Comparing the two years in this dataset, one cannot only see a decrease in tweets (from 1.311.190 in 2020 to 1.199.558 in 2021), but also a decrease in usage of shared locations (from 2,71% to 1,86%).

2.3 Usage of User Profile Location in Tweets

User profile location is a field in the Twitter user profile that can be entered during registration or even later. It is a optional text field where any text can be entered and not a selection of existing cities can be made. However, localizing using the user profile location, can be a good alternative to others like using shared locations [LRK16].

In the given dataset 2.039.120 tweets out of 2.510.748 (81,22%) have something specified in the user profile location. Since this can be anything and do not have to be meaningful, it is more interesting to look at the numbers of certain cities or countries. Using a dataset of cities in English language from around the world from Kaggle[1], a comparison can be made with the user profile location information. In 230.238 (9,17%) of the tweets exactly a city from the dataset was specified (*user profile location = dataset.city*). In 206.160 (8,21%) of the tweets exactly a country from the dataset was specified (*user profile location = dataset.country*). Exact matches of European capitals are found in 68.956 (2,75%) and exact matches of European countries in 55.170 (2,20%) of the tweets.

City	Tweets 2020	Tweets 2021	Total Tweets
London	60.530	57.092	117.622
Brussels	11.488	11.302	22.790
Berlin	10.211	8.270	18.481
Paris	5.939	5.917	11.856
Dublin	3.664	4.148	7.812
Madrid	2.643	1.802	4.445
Amsterdam	2.203	2.100	4.303
Helsinki	2.238	1.951	4.189
Vienna	1.611	1.739	3.350
Rome	1.672	1.645	3.317
Stockholm	1.600	1.535	3.135
Copenhagen	1.313	1.060	2.373
Luxembourg	1.027	964	1.991
Athens	926	980	1.906
Oslo	788	719	1.507
Bern	383	422	805
Prague	461	316	777
Lisbon	411	306	717
Warsaw	337	298	635
Budapest	282	292	574

Table 2.3: Top 20 amount of tweets per European capitals from the Business Understanding phase dataset

[1] https://www.kaggle.com/datasets/max-mind/world-cities-database (Last accessed on 2022-07-06)

In the example of Vienna (Austria), 521 user profile locations of tweets match exactly the city name 'Vienna' and 328 the country name 'Austria'. However, in the case of Vienna several reasonable expressions are possible, such as 'Vienna', 'Wien', 'Vienna, Austria' and surely some more. Since so far only the exact match was counted a lot of improvement is possible here. In a first step the search for an exact match is replaced by the city surrounded with wildcards (*user profile location = %dataset.city%*). This increases the number up to 3.350 tweets with 'Vienna' in its user profile location. Table 2.3 shows European capitals with the 20 most occurrences based on the search with wildcards. With this approach 216.311 (8,62%) tweets can be matched to European capitals.

An additional way to increase the amount of data is to add the area around of cities. This is possible with the use of longitude and latitude. The easiest way is to start from a capital city +- a certain value in longitude and latitude, which then results in a square perimeter. Using the example of Vienna and the value of 0,15, 54 more tweets can be assigned to Vienna. The additional cities found around Vienna are 'Korneuburg', 'Klosterneuburg', 'Wiener Neudorf' and 'Brunn am Gebirge' and are from six different users.

2.4 Conclusion and Data Mining Goals

Twitter is a good data source for analysis because of the amount of data available. If one understands how to deal with the challenges, one can cleanly prepare data so that it is usable for processing and analysis. With the conditions selected at the beginning, valuable conclusions can be drawn. The questions on possible hashtags, sharing locations and user profile location in tweets are examined in this phase in order to better understand the existing data on Twitter. This allows a data set to be created for further analysis in order to answer the research questions.

After reviewing the occurring hashtags related to climate change, the following hashtags (9) are selected as search parameters: #climatechange, #globalwarming, #climatecrisis, #climateaction, #climateemergency, #climatechangeisreal, #climatejustice, #climatebrawl and #climatestrike. The hashtag #climate is not selected, as are others (e.g. #covid19, #environment or #sustainability) due to insufficient significance to climate change [HZ15]. The choice is made for more general hashtags, rather than specific topics such as drought or sea level rise, so as not to bias the results too much with such hashtags. Since tweets with the hashtag #cop26 may be about the conference itself and not necessarily about climate change, as well as possibly expressing information about results and not at all thoughts of users, this hashtag is not selected as search parameter. Also the hashtag #climatehoax is not chosen as search parameter, because it represents disbelievers [TBCS20] and also can bias the results. These deselected hashtags as search parameters will appear again in the new dataset, but then they will be used with a hashtag related to climate change.

Since the research questions deal with European cities, comparing sharing locations and user profile locations on these data is necessary. In the approach searching tweets within user profile location using wildcards 216.311 out of 2.510.748 (8,62%) tweets can be matched to European capitals. In comparison, only 2.380 (0,09%) tweets are gathered from European capitals by using sharing location. Thus, the amount of tweets with user profile location is significantly higher than tweets with sharing location. Unfortunately searching for tweets within user profile location using wildcards is not quite accurate enough and can lead to certain inaccuracies, but is still a good indicator in the amount of tweets per cities found. Further improvements need to be made in the Data Preparation phase.

When searching with user profile location for nearby cities using longitude and latitude, the risk is too high for the small proportion of tweets added and so this will not be done in this work. If several cities are added to the search, there are also added sources of errors. For example, city names can occur several times in the country/world or city names can be part of other city names ('Wiener Neudorf' contains 'Wien').

Therefore, the hashtags mentioned above are used as search parameters and the user profile location is used as a criterion for assigning tweets to cities in order to be able to answer the research questions. The research questions deal with two different topics in the field of text mining. First, to find out what the mood is and second, which words are polarizing. For the analysis of mood, sentiment analysis can be used to find the most negative cities with respect to the mood and to find out if there are similarities between coastal cities and inland cities. For the analysis of polarizing words it is necessary to make comparisons with methods that allow to include the ratio, so that comparisons of cities can be made. But also absolute methods such as word clouds or comparisons of the top words allow an evaluation of the similarities of cities.

CHAPTER 3

Data Understanding

After the first insights into Twitter have been delivered in the Business Understanding phase, it is now a matter of deepening these insights and looking at concrete cases in the dataset. To do this, a script is needed that extracts the data with the predefined parameters. In order to better understand the results, the data is described in further detail. Then, data is loaded into a database and deeper analysis is done, which brings new insights from the data and identifies data quality problems. These problems must then be resolved in the next phase (Data Preparation). In this phase not only tweets are gathered and analyzed, but also a city dataset with European capitals containing information to help answering the research questions.

3.1 Collection of Data

There are different approaches to collect data. One can use APIs or libraries from various programming languages to retrieve data by writing a script in a selected programming language. Another option is to use already existing datasets. In recent years, there has been more and more publicly available data [BCN20], which can be found via certain dataset searches, such as Google Dataset Search[1]. In addition there is still the option to build a custom dataset.

3.1.1 Twitter Dataset

To collect a dataset of tweets, the option was chosen to write a custom script that extracts the data and stores it in a .csv file. The script was written in the Python programming language using snscrape[2], which is a scraper for social networking services. It allows to pull tweets with various filters, such as keywords, language or time period.

[1] https://datasetsearch.research.google.com/ (Last accessed on 2022-07-06)
[2] https://github.com/JustAnotherArchivist/snscrape (Last accessed on 2022-07-06)

Therefore the following code snippet shows the defined parameters from the Business Understanding phase for the search of tweets within the script:

```
keywords = ’#climatechange OR #globalwarming OR #climatecrisis
OR #climateaction OR #climateemergency OR #climatechangeisreal
OR #climatejustice OR #climatebrawl OR #climatestrike’
lang = ’en’
since = ’2020-01-01’
until = ’2022-01-01’

tweets = sntwitter.TwitterSearchScraper(
            keywords
            + ’ lang:’ + lang
            + ’ since:’ + since
            + ’ until:’ + until
        ).get_items()
```

In the search multiple keywords can be specified with the separation ’OR’. This way single occurrences of hashtags or also combinations of hashtags are found. The result of the search is stored first in a list of tweets, second converted to a dataframe and then finally written into a .csv file.

3.1.2 City Dataset

Since there was no dataset that met the requirements for the analyses in this work, a custom dataset was created. This is because it needs a list of European capitals both in English and in the official language of the country. It is necessary because some Twitter users do not provide their origin in English, but in their own language.

For this purpose, all European capitals were listed in English and additionally, if available, the cities were also added in official language. In order to check whether the declaration of the cities in English language was listed correctly, it was compared with the data set of Kaggle with cities in English language. This verified that the list did not contain any typos or spelling errors.

Additionally, meta-information about the cities was added, such as the height above sea level and distance from the coast. For the height above sea level information from Wikipedia was used. The distance of cities from the coast was measured using Google Maps starting from the city center to the nearest sea access. If the city is located directly on the sea or outlets of the sea, 0km was given as the distance.

3.2 Description of Data

The prepared .csv files from the previous step are then loaded into a database to enable analysis with SQL. These datasets contain several columns each with specific information.

In order to be able to make analyses with it, it is important first to understand what kind of data is available at all and what the individual columns can contain for characteristics. Therefore the data and it attribute types are described below.

3.2.1 Twitter Dataset

The dataset of the tweets consists of standard information that can be extracted from Twitter. Some of these are Twitter-specific terms and are not described here, but information about it can be found on Twitter Help Center[3]. The dataset contains the following fields:

TweetId (varchar) Contains the unique tweet ID.

PostingTime (datetime) Indicates the creation date and time of the tweet.

Text (varchar) Contains the text of the tweet including all hashtags used.

Hashtags (varchar) Includes all hashtags used in the tweet.

Language (varchar) Indicates the language of the tweet.

City (varchar) If the location information is shared in a tweet, the city is available in this field.

Country (varchar) If the location information is shared in a tweet, the country is available in this field.

Longitude (varchar) If the location information is shared in a tweet, the longitude is available in this field.

Latitude (varchar) If the location information is shared in a tweet, the latitude is available in this field.

Replies (int) Indicates the number of replies of the tweet.

Retweets (int) Indicates the number of retweets of the tweet.

Likes (int) Indicates the number of likes of the tweet.

Quotes (int) Indicates the number of quotes of the tweet.

Username (varchar) Contains the username of the user who sent the tweet.

Location (varchar) Contains the user profile location, if specified, of the user who sent the tweet.

Description (varchar) Contains the description, if specified, of the user who sent the tweet.

[3] https://help.twitter.com/en (Last accessed on 2022-07-06)

UserCreated (datetime) Indicates the creation date of the account of the user who sent the tweet.

Followers (int) Indicates the number of followers of the user who sent the tweet.

Friends (int) Indicates the number of followed users of the user who sent the tweet.

Staties (int) Indicates the number of staties of the user who sent the tweet.

Favourites (int) Indicates the number of favourite tweets of the user who sent the tweet.

3.2.2 City Dataset

In order to subsequently assign the tweets to cities, a separate city dataset is required. It consists of the 47 European capitals[4], their countries and additional meta information about the cities. This dataset contains the following fields:

City (varchar) Is the name of the city. This field contains the official and English language versions of city names, as this column will be used to find European cities in user profile locations. For example, for Vienna 'Wien' and 'Vienna' are included here.

City_eng (varchar) Is the name of the city in English language. This field is to harmonize the different versions of city names in the field City using the English name of it.

Country (varchar) Is the country in which the city is located in English language.

Elevation (int) This field indicates the elevation above sea level of the city in meters.

Coast_Distance (int) This field indicates the distance of the city to the nearest coast in kilometers.

In table 3.1 the example of Vienna from this dataset can be found. It shows the two different values of the 'City' field for Vienna, one in the official language and one in English language.

City	City_eng	Country	Elevation	Coast_Distance
Vienna	Vienna	Austria	170	340
Wien	Vienna	Austria	170	340

Table 3.1: City dataset of Vienna

[4] https://www.laender-lexikon.de/Europa_L%C3%A4nder_A-Z (Last accessed on 2022-07-06)

3.3 Exploration of Data

After the first analyses in the Business Understanding phase to understand Twitter in general, the aim here is to better understand the dataset and identify possible data quality problems. Furthermore, insights into the self-created dataset for cities are given.

3.3.1 Twitter Dataset

Based on the findings and the resulted new search parameters from the Business Understanding phase, a new dataset was extracted. This dataset will serve as the basis for all further analyses. In order to present the findings from this dataset, comparisons are made with the dataset from the Business Understanding phase in such a way to highlight the changes resulting from the adjustment of the search parameters. In addition, further in-depth analyses are also performed to show more details from the dataset. In the following, the dataset from the Business Understanding phase is referred to as the old dataset and the dataset from the Data Understanding phase is referred to as the new dataset or simply the dataset.

With the new selected hashtags 3.768.425 tweets were extracted. In comparison, 2.510.748 were extracted with the initial selection of hashtags. That is an increase in tweets of quite exactly 50%. In the new dataset, unlike the old dataset, there is an increase in tweets from 2020 to 2021. The largest volume of tweets can be found around the time of COP26 (October and November 2021) and early 2020 (January and February) before Covid19 broke out, see table 3.2.

Month	**Year**	**Tweets**
January	2020	286.331
November	2021	241.727
October	2021	212.465
February	2020	200.418

Table 3.2: Months with the most tweets

When checking the data for NULL values, it turns out that, with a few logical exceptions, no NULL values were found. The exceptions concern optional fields like GPS data, user profile location or user description. Only the hashtags field, where one would expect all records to be set when searching for hashtags, contains NULL values (314 rows). This occurs when Twitter blacks out the user's content due to copyright violations or violations of the Twitter Media Policy. Another reason why the field may be NULL is when hashtags are used without whitespaces between the hashtags. Apparently the interface cannot handle this and thus does not recognize hashtags. The search looks for the hashtags in the text field, so tweets that contain hashtags concatenated together can still be identified and subtracted.

Hashtags

Due to the higher number of tweets overall, the number of hashtags has also increased compared to the old dataset (from 392.578 to 498.312 different hashtags). On average, 3,72 hashtags are used and the tweet with the most hashtags has 39 hashtags.

In table 3.3 the number of tweets of already discussed hashtags from the Business Understanding phase can be seen. The table shows where the hashtags are in a ranking by the number of occurrences, how often the hashtags occur, and what the difference is to the old dataset.

#	Hashtag	Occurrences	Difference
1	#climatechange	1.811.088	-2.768
2	#climatecrisis	892.281	+687.982
3	#climateaction	863.801	+668.023
4	#climateemergency	597.574	+486.773
5	#globalwarming	217.858	-1.381
6	#cop26	208.712	+111.568
7	#climate	155.122	-521.744
8	#environment	134.930	+8.424
9	#sustainability	134.851	+21.036
10	#climatejustice	108.268	+88.930
11	#covid19	93.109	+27.375
13	#climatechangeisreal	67.132	+43.798
18	#climatestrike	47.009	+33.642
28	#climatebrawl	33.172	+13.882
150	#climatehoax	7.680	+729

Table 3.3: Occurences of hashtags in the new dataset

Interestingly, the number of the hashtags #climatechage and #globalwarming is different from the old dataset, although again both were defined as search parameters. An analysis of tweets showed that this could be due to user settings, among other things. Some users are present in the old dataset, in the new dataset all tweets of these users were missing. Thus, there may have been a change in the user profile settings in the time between the subtraction of the two datasets, which could explain the difference.

The other hashtags that were used as new search parameters logically record an increase in occurrences. Hence, the hashtags #climatecrisis, #climateaction, and #climateemergency were able to increase the number significantly. The only difference in the top 10 in occurrences compared to the old dataset is the hashtag #climatejustice, which replaced the hashtag #covid19. The other hashtags that were used as new search parameters are apparently not as widespread on Twitter and do not produce a large increasing number.

The hashtag #cop26 apparently also occurs quite often together with the other newly added hashtags and recorded a large increase in occurrences. The hashtags #environment and #sustainability also managed to stay in the top 10. The hashtag #covid19 just slipped out of the top 10, landing on the 11th rank. These mentioned hashtags saw an increase in occurrences, as they appear to be often tweeted in relation to climate change and in the new dataset additional climate change related hashtags were added. Only the hashtag #climate, which was assumed not to be as relevant to climate change and so was removed as search parameter, has a large reduction. Though this hashtag still managed to stay in the top 10. This shows that this hashtag seems to have been a good choice in the initial dataset to find related hashtags to climate change. Last, the hashtag #climatehoax, which represents more disbelievers [TBCS20], did not see a very large increase in the number of tweets.

User Profile Location

In the new dataset, a user profile location is specified in 3.044.500 of the tweets. This represents a share of 80,79% (old dataset = 81,22%). Using a wildcard search (*user profile location = %Cities.City_eng%*), it was possible to find 7,86% of the tweets that can be assigned to a European capitals (old dataset = 8,62%).

City	Tweets	Difference
London	172.026	+54.404
Brussels	28.624	+5.834
Berlin	19.558	+1.077
Dublin	13.552	+5.740
Paris	12.502	+646
Amsterdam	5.667	+1.364
Stockholm	4.643	+1.508
Vienna	4.583	+1.233
Rome	4.553	+1.236
Helsinki	4.511	+322
Madrid	4.193	-252
Copenhagen	3.161	+788
Athens	2.833	+927
Luxembourg	2.436	+445
Oslo	2.325	+818
Bern	1.326	+521
Lisbon	1.253	+536
Bucharest	907	+693
Warsaw	848	+213
Prague	803	+26

Table 3.4: Top 20 amount of tweets per European capitals from the Data Understanding phase dataset

Although the relative numbers have fallen, the absolute numbers have risen due to the volume of data. In the end, the absolute numbers are decisive in determining whether meaningful statements can be made. The results should still be treated with caution, as they still contain inaccuracies that will be corrected in the Data Preparation phase.

Table 3.4 shows the top 20 occurrences of European capitals with the wildcard search in the new dataset. This table states the number of tweets and the difference of the amount from the old dataset. For the cities in this list, there has been only one change from the top 20 capitals in table 2.3 in the old dataset: Bucharest is included instead of Budapest. Bucharest thus also records the highest increase percentage-wise of all cities with over three times the amount of tweets in the new dataset. In contrast, Madrid records a lower data volume of tweets compared to the old dataset. This is mainly due to the removal of the hashtag #climate as a search parameter. The old data set contained 679 tweets with this hashtag and the new data set only contains 72 occurrences in Madrid. This shows that the hashtag #climate was often not used in connection with the identified climate change hashtags in Madrid. All other cities record no unusual changes in the dataset, only slight fluctuations in the magnitude of the change.

Users

It is also interesting to look at the users of the corresponding dataset. The whole dataset contains 591.114 different users, of which 41.998 (7,1%) can be assigned to European capitals using the wildcard search. Table 3.5 shows the ten cities from the top 20 European capitals with the highest tweet per user share. The number of posts per user in the top 20 European capitals ranges from 4,61 to 11,75, with a mean of 6,41.

City	Tweets per User	Different Users	Tweets
Brussels	11,75	2.436	28.624
Bucharest	9,45	96	907
Rome	7,73	589	4.553
Luxembourg	7,52	324	2.436
London	7,19	23.942	172.026
Berlin	7,10	2755	19.558
Vienna	7,06	649	4.583
Athens	7,02	404	2.833
Stockholm	5,73	810	4.643
Warsaw	5,62	151	848

Table 3.5: Top 10 tweets per User

If one looks at the users with the most tweets, see table 3.6, one sees that they belong to the two cities that also have the most tweets, see 3.4. The user with the username *chriscartw83* has the most tweets in European capitals, with the amount of 9.294. From the user profile description it can be seen that this user tracked data related to events (*'[..] I track related data & events [..]'*). In this user's tweets, several start with the word 'daily' and are followed by recorded data. However, the account is not a pure bot for tracking data, because there are also many self-written tweets to be found. For the other users in the table, such characteristics could not be found in their description or at a first sight in their tweets. All these users work or are activists in the field of climate change, which explains their high number of tweets.

Tweets	Username	City	User Profile Description
9.294	chriscartw83	London	Mankind will not lift a finger to prevent the climate crisis . I track related data & events Empty pledges and lies were the best we could do Tragedy is next.
7.597	sustainableuni1	London	Writer in London sustain-blog.com | #climatechange #sustainability
6.149	AlexWitzleben	Brussels	Retired @EU_Commission PhD CAU Kiel #EUI #EconomicHistory #EUinfluencer 2021 second in #sustainability #ClimateActionNow #EUGreenDeal #FutureOfEurope #startups
1.623	tom_burke_47	London	Environmentalist. Chairman & Founding Director of @E3G. Particularly active on energy & climate issues. Prominent critic of government nuclear power policy.
1.463	Redjont	London	Retired, Climate Activist, Anti Racist, Live Music Lover and Grandfather. HeHim #RebelForLife #FreedomToProtest #AntiRacism #SocialJustice #ClimateJustice

Table 3.6: Top 5 users per amount of tweets

Texts

The most important part of the dataset for text analyses are the tweets themselves, which are in the field called Text. Although there is a separate Hashtags field, all hashtags are included in the Text field as well, since they can appear anywhere in the tweet. In addition, other characteristics such as tagging users using @ and the username after it (e.g. @johndoe) or hyperlinks are present in the tweets. An analysis of the dataset shows that these two characteristics occur very frequently in tweets. The @ character occurs in 1.759.700 of the tweets, which represents a percentage of 46,70% of the tweets. Since @ is also used in email addresses, there will be fewer tweets with marks of users. Though, the number of tweets will not be much lower, as only 585 occurrences from the three most used domains[5] (gmail.com, yahoo.com and hotmail.com) are present in this dataset.

[5] https://email-verify.my-addr.com/list-of-most-popular-email-domains.php (Last accessed on 2022-07-07)

The use of hyperlinks is even more prevalent, occurring in 3.000.432 tweets (79,62%). However, a distinction must also be made here between different hyperlinks. Twitter does not only display external references as hyperlinks in the data, but also shared images, videos or quotes from other tweets.

When checking for NULL values in all fields, it has already been noticed that there are tweets that have been blacked out by Twitter. This happens in case of violations of copyright or violations of Twitter Media Policy. Such tweets then contain the following texts in the dataset:

- This Tweet from @Username has been withheld in response to a report from the copyright holder. Learn more.
- @Username's account is temporarily unavailable because it violates the Twitter Media Policy. Learn more.

But the number of them in the dataset is very small. There are ten tweets with violations of the copyright and 49 tweets with violations of Twitter Media Policy. With violations of Twitter Media Policy all tweets of a user are blacked out, whereas with violations of copyright only single tweets are blacked out.

In other analyses and the screening of tweets, it was possible to find some tweets that are similar and belong to the same user. This could be, for example, users who write similar tweets and tag different users. More often, however, this is the case with bots or users who share similar information at regular intervals. For example, the user with the most tweets in this dataset tracks data on a daily basis and thus some of his tweets have similarities.

3.3.2 City Dataset

The city dataset contains 47 different cities, namely the capitals of Europe. With the different names in English language, official language or common use of the cities, this dataset comes to a total of 77 city names. This is because 23 of the 47 cities have at least a second variation of the name (e.g. 'Wien' and 'Vienna' for Vienna). The city with the most different variations is Brussels, namely four: Brussels, Brussel, Brüssel, Bruxelles. This is due to the fact that the European Parliament has its headquarters there and the different languages that are spoken in Brussels.

Among these 47 European capitals, 13 are located directly on the sea or outlets of the sea, as listed in table 3.7. Comparing this table with the table showing the top 20 amount of tweets per European capitals in the new dataset (see table 3.4), one sees that eight of the 20 cities listed there are located on the coast.

Valletta and Monaco are such capitals that are located on the sea and special is the fact that the elevations are 54m (Valletta) and 65m (Monaco) above sea level, as the centers are located significantly higher than sea level. From the capitals that are not located by

City_eng	Country	Elevation	Coast_Distance
Amsterdam	Netherlands	-2	0
Athens	Greece	20	0
Copenhagen	Denmark	5	0
Dublin	Ireland	8	0
Helsinki	Finland	25	0
Lisbon	Portugal	15	0
Monaco	Monaco	65	0
Oslo	Norway	12	0
Reykjavik	Iceland	15	0
Riga	Latvia	8	0
Stockholm	Sweden	15	0
Tallinn	Estonia	9	0
Valletta	Malta	54	0

Table 3.7: Cities on the coast

the sea, Rome and London have the lowest elevation above sea level (14m) and are also fairly close to the sea with 25km and 50km. Also noticeable when looking at the numbers is San Marino, as it is only 20km from the sea away, but has an elevation above sea level of 749m, as it is on a mountain. The overall highest capital of Europe is Andorra la Vella with 1.023m above sea level.

3.4 Conclusion and Identification of Data Quality Problems

Very useful data has already been collected, both from Twitter and for the city dataset. The selected search parameters from the Business Understanding phase were able to increase the amount of data in the Twitter dataset, and when evaluating the hashtags, most of these hashtags also increased significantly in the ranking by occurrence. Some hashtags, however, are simply not used that often on Twitter and consequently could not show the expected effect of significantly more tweets. In contrast, some hashtags that were not selected as search parameters occurred quite often in the dataset, as they appear to be often tweeted in relation to climate change. Overall, the dataset has achieved a sufficient amount of data that can be attributed to climate change and thus analyzed.

Looking at the most important fields, one can see that there are no NULL values in the data except for the Hashtags field. Surprisingly, there are no values in this field for some tweets because Twitter has blacked them out or hashtags were concatenated together and the interface could not handle that. Anyway, this field is not relevant for the analyses, because the information from the field Text is used, so it does not need any further processing.

By analyzing users and tweets, a potential improvement of the dataset could be identified in the text field, in addition to standard approaches for data preprocessing in the area of text mining (filtering and lemmatization) [APA+17]. It is necessary to filter out hashtags, tagging and hyperlinks, as well as identify and remove similar tweets and data tracking from the dataset. Blacked out tweets can be filtered out easily because the whole tweet with such text phrases can be ignored when extracting the dataset.

By using the user profile location, assignments to European capitals could already be made, but they were not precise enough. Since this is a free text field, the users can enter all possible variations of cities here. In addition, there is the problem that some city names might occur in other city or country names and thus cause inaccuracies. Therefore it will be necessary to do a very precise filtering including exclusion of errors when joining the Twitter and city dataset. The city dataset itself is a small and self-created dataset that does not show any data quality issues and is already suitable for analysis.

CHAPTER 4

Data Preparation

This phase is about preparing the data so that it can be analyzed in the next phases. Both datasets (Twitter and City) are still separate tables that have to be joined together. In the previous phases this was done with an initial approach, namely a wildcard search with the English city names. However, this search is not accurate enough and contains incorrect mappings of tweets to cities. Therefore, these two datasets are to be joined with an approach that is as accurate as possible. Furthermore, several text preparation steps are performed to prepare the tweets for text mining techniques, so that at the end of this phase a final dataset is available for the Modeling phase. These steps include cleaning up the text itself by removing hashtags or hyperlinks, as well as removing similar tweets.

4.1 Assign Tweets to European Capitals

There are only a few approaches in the literature that try to make an association by using user profile location as source of information. Most of them use GPS data, which, however, does not provide representative data for Twitter [KKP+21]. If one can find methods to assign tweets to a location based on text or user profile location, then there is potential in this and would be an essential alternative of using GPS data [LRK16]. Alex et al [ALG+16], for example, created a geoparser for the user profile location field with which they were able to achieve 90% accuracy. But also other fields or information can be used to try to make a mapping [ZHS18], such as using tweet texts and the friendship networks of users [RAP+13]. In this research, an attempt is made to assign tweets to cities based on characteristics found in the field user profile location. For this purpose, an intensive analysis of the characteristics of user profile locations was made, the problems found were shown and a solution was implemented as a result.

4.1.1 Approaches, Challenges and Solution of the Assignment

Up to now, the searches for cities in the user profile location were always made with the English names of the capital cities, in order to avoid additional errors with other variations of the cities in the initial analyses. However, the goal is to use these as well, because there are other valid spellings for some cities. Therefore this is extended now also with the assigning of city names in official language.

In order to prevent that the assignment of cities also returns matches for other expressions of the term in the user profile location field, the wildcard search was adapted in a first approach. So the search was done exactly for the city name and also whitespaces(’ ’) and commas(’,’) were allowed before and after it, as can be seen in the code snippet:

```
tweets.location like ’%’ + cities.city +  ’ %’
or tweets.location like ’% ’ + cities.city + ’%’
or tweets.location like ’%’ + cities.city + ’,%’
or tweets.location like ’%, ’ + cities.city + ’%’
or t.location =  c.city
```

This should cover the most common ways to indicate a location. Allowed examples for Vienna would be ’Vienna’, ’Vienna, Austria’, ’Vienna Austria’, ’Austria Vienna’, ’Austria, Vienna’ among others with the German names. In reviewing and comparing this approach to the previous wildcard search (*tweets.location = ’%cities.city%’*), other possible occurrences could be identified. There are sometimes several cities specified and thus enumerations are used. Therefore the next approach is to allow the different enumeration symbols as additional options. These are such symbols as ’/’, ’-’, ’&’, ’#’, which have to be allowed before or after a city name. Unfortunately, further additional special symbols are sometimes used or cities can be strung together without spaces or symbols. This makes such a logic very complicated.

This leads back to the original approach of assigning cities, no matter what characters appear before or after them. With the 47 European capitals, this results in 8.747 different entries of user profile locations, whereby some entries occur more than once, since they can be assigned to various cities. But this is not the only problem that can appear, because during a manual review many more challenges and wrong assignments were identified. The longer the information was (e.g. by specifying postal code, district or enumerations), the easier it was to confirm the correctness. For some entries, it was necessary to do research to confirm the assignment. It can be assumed that for cities without a more detailed description, such as the country, the capitals in Europe are meant, since the other possible cities with the same name are not as well known and require further descriptions. The following findings were obtained from the data:

- Sometimes the city is specified in several languages and is then found multiple times in the city dataset with the different variations of the city, for example if ’Vienna/Wien’ is stated.

- Sometimes cities are specified in an enumeration and are thus assigned to multiple cities of the city dataset.
- Sometimes the areas/districts of the cities are given in addition to the city name
- Comma-separated names can be enumerations or an addition to the city name, such as specifying a country or state in the USA.
- Some European capitals (such as Athens, London, Dublin, Vienna or Berlin among others) also appear several times in the USA or Canada and are marked with a state of theses countries.
- Bern, Brussels and San Marino occur in other variations of themselves (e.g. ’Bern’ in ’Berne’)
- Some cities (e.g. Bern, Riga or Rome) occur in various terms which cannot be assigned to the respective city. For example, ’Roma’ occurs in ’Romania’, which leads to the assignment of Romanian cities to Rome.
- Some cities appear in street names, which were sometimes also given in the location.
- In Ireland there is not only the city of Dublin, but also a county called Dublin.
- Luxembourg is not only the city name, but also the name of the country, which is why there are a few cities from Luxembourg that state it as a country identifier in addition to the city name.

However, all these findings can be resolved so that they do not appear as problems in the dataset. There are often entries for multiple cities in the user profile location and therefore an exclusion of these tweets is not made. Thus, these tweets may be valid for multiple European cities. If tweets are assigned to a particular city multiple times, these duplicates are filtered out so that each tweet per city only occurs once. There are also findings where terms or cities are assigned but are obviously not correct. Some of the findings contain examples of this, and some findings (e.g. country or state in the USA as additional identifier) also served as a support for identifying incorrect assignments. Therefore for each city identifiers were defined that are not valid for this city and thus should not be assigned to this city.

```
(city_eng like ’%Athens%’
and location not like ’%Athens, Ohio%’ [..])
or (city_eng like ’%Dublin%’
and location not like ’%County Dublin%’
and location not like ’%Co Dublin%’
and location not like ’%Co. Dublin%’[..])
or (city_eng like ’%Rome%’
and location not like ’%Romania%’ [..])
```

In this code snippet, three different findings are shown as examples. In the case of Athens, an additional designation to the city is given, which makes it clear that it does not refer to Athens in Greece. For Dublin different variations have been excluded that mean County Dublin in Ireland and not the city of Dublin. And third, the example of Rome is shown not to allow a occurrence of 'Romania'. For some cities such problems did not exist, because the name is either unique or simply does not occur in other names. For the other cities, such exclusions were defined so that they were not assigned incorrectly. Thus, a very accurate assignment to European capitals was made.

4.1.2 Exploration of Data after the Assignment

After the assignment of the tweets to European capitals has been made, it was possible to take another look at the number of tweets that were assigned after joining. So far, the two data sets (Tweets and City) have been reviewed individually or with approaches to join them. In the table 4.1 the top 20 European capitals are again listed based on the number of tweets. The column 'Difference' shows the difference from the evaluation of the dataset in the Data Understanding phase and the column 'By the Sea?' indicates whether the city is by the sea or not.

City	Tweets	Difference	By the Sea?
London	170.776	-1.250	no
Brussels	30.508	+1.884	no
Berlin	19.472	-86	no
Dublin	13.215	-337	yes
Paris	12.455	-47	no
Amsterdam	5.648	-19	yes
Rome	5.509	+956	no
Vienna	5.324	+741	no
Stockholm	4.643	0	yes
Helsinki	4.567	+56	yes
Madrid	4.192	-1	no
Copenhagen	3.584	+423	yes
Athens	2.569	-264	yes
Luxembourg	2.348	-88	no
Oslo	2.322	-3	yes
Lisbon	1.622	+369	yes
Warsaw	1.124	+276	no
Bucharest	934	+27	no
Prague	870	+67	no
Bern	667	-659	no

Table 4.1: Top 20 amount of tweets per European capitals after city assignment

This table shows very clearly the effects of the two changes that were made compared to previous evaluations. First, when assigning tweets to European capitals, not only the English names were used, but also the city names in the official language. These cities were thus able to register an increase in tweets. Examples include Brussels, Rome, or Vienna, among others. The second change in the assignment was excluded terms, which lead to fewer tweets for those cities. Although terms were also excluded for Brussels, Rome and Vienna, the assignment of cities in official language shows more effect. However, there are also cities that have no other name. These show a decrease in tweets (e.g. London, Berlin or Dublin). Athens (due to its occurrences in the USA) and Bern are most affected by this excluded terms. Bern is a short name and is a term that occurs often in other cities or designations. Stockholm is an example of a city where there is no other name, nor was it necessary to exclude terms. Outside the top 20 there are other such examples.

To make a comparison between coastal and inland cities, it is also interesting to see how many of the top 20 amount of tweets per European capitals are by the sea and how many are not. Currently in the top 20 a total of eight cities are by the sea and twelve are not, so it's already quite balanced. If one looks at the top 16 of them, it's balanced with eight cities by the sea and eight cities inland and all of these cities have over 1500 tweets.

4.2 Text Preparation

In order to be able to draw meaningful conclusions from the tweets, the tweet texts need to be prepared. That is why this preparation is purely about the tweet texts. If one takes the raw tweets, some disturbing factors are still contained, such as hyperlinks. Such factors have to be removed in a text cleaning. But not only in the tweets themselves there are issues, also users can be such. Some users can be bots and post very similar things every day. These can be tracking data about the weather, but also advertisements for certain events or other things. Such tweets must be kept to a minimum.

4.2.1 Text Cleaning

Text cleaning was about removing Twitter characteristics, unnecessary symbols and the stop words, so that no disturbing factors were present in the analysis of the texts. Two approaches were used to accomplish this. On the one hand, regular expression (short regex) was used to find parts of the text that were not considered suitable for text analysis (e.g. hashtags). On the other hand, the Natural Language Toolkit (NLTK)[1] was used. NLTK is a platform for building Python programs to work with human language data. It is an open source Python library for NLP. This was used to remove stop words and to perform lemmatization.

In order to remove inappropriate text passages, they were searched for with regex and replaced with a blank text ", except for symbols, which were replaced with a space ' '. To find appropriate regex conditions, they were tested with several examples and thus

[1] https://www.nltk.org/(Last accessed on 2022-08-24)

optimized. In the following the applied removals and word cleanups are described in the order of execution in the code.

Remove hashtags: With hashtags (#), there are special features to consider when using them on Twitter and thus when removing them. Hashtags begin with the symbol # and are followed by a string of letters and numbers, with at least one letter. Special characters (symbols or spaces) cannot be used within hashtags as they would end the hashtag, except in the case of an underscore ('#it's' results in the hashtag #it and #under_score remains #under_score). Thus, the regular expression #[\w]* was used to find hashtags. Since symbols and numbers are also removed later, there is no need to make sure that at least one letter is included in this regex. A Python code example of the function to remove hashtags can be seen in the following.

```
def removeHashtags(text):
    return re.sub(r'#\w*', '', text)
```

Remove emails: The @ symbol is known in Twitter usage as a symbol to mark users. In common usage, however, it is usually associated with emails. Since emails have text before and after the @ and tags only after the @, emails need to be cleaned from tweets before tags. Emails allow to have a combination of letters, numbers and certain symbols before the @. After the @ is the domain, which is usually very similar in structure, resulting in the regular expression [A-Za-z0-9.!#$%&’*+\-\/=? ^_‘{|}~]+@[A-Za-z]+\.[A-Za-z]{2,3} to filter emails from tweets.

Remove user tags: The tagging of users is similar to the use of hashtags. To tag users, the @ symbol is used in front of usernames to mark them. For the selection of usernames during registration, Twitter has rules[2] on how to create them. According to Twitter, usernames must contain alphanumeric characters (letters [A-Za-z] and numbers [0-9]), additionally only underscores are allowed. Also, usernames cannot contain the words 'Twitter' or 'Admin' unless it is an official Twitter account. This leads to @[\w]* as a regular expression for searching user tags, since emails have already been filtered out.

Remove hyperlinks: On Twitter, hyperlinks are not only external references, but also internal ones, such as shared images. This simplifies the removal of such content, since hyperlinks are always structured very similarly. At the beginning there is either https, http or www followed by the website or a shortened URL for internal references. With the regular expression www\.[^\s]+)|(https?://[^\s]+ hyperlinks can be found in the tweets.

Remove conversion errors: When using different data storages it can happen that some symbols are displayed differently when converting. For example, reserved characters in HTML are replaced with character entities [3]. This causes the string &amp; to appear

[2] https://help.twitter.com/en/managing-your-account/twitter-username-rules(Last accessed on 2022-08-24)

[3] https://www.freeformatter.com/html-entities.html(Last accessed on 2022-08-24)

in the text instead of the symbol &. A similar situation occurs with < (&lt;) and > (&gt;). These can be easily found and filtered out with the regular expression *&amp|&lt;|&gt;*.

Remove symbols and numbers: After the specifics, such as hashtags, emails, tags, hyperlinks and conversion errors have been removed, all other symbols and numbers can now also be identified and deleted using the regular expression *[^A-Za-z\s]*, since it does not need them in text analyses anymore. This leaves only letters and spaces.

Lemmatization: Since words can be used in different forms or tenses, it requires lexicon normalization to unify words with the same origin. Two approaches to this are stemming and lemmatization. Stemming reduces the words to their root/stem by looking at the suffix[APA$^+$17, SKMM20]. Lemmatization is looking for the lemma which is a root word and not a root stem[SKMM20]. The main difference is, that the output from stemming is not necessarily a real word from the lexicon, however, this is the case with lemmatization[SKMM20]. An example is the word studies, which returns studi in stemming and study[SKMM20] in lemmatization. It can be concluded that the advantage of lemmatization is that it produces actual words as a result. The disadvantage, however, is that this requires further processing in order to produce meaningful results. The so called Part-of-Speech(POS) tagging is needed to find out the grammatical group (noun, verb, ...) of a word, which is done based on the context of the word in a sentence [SM19]. NLTK offers modules for implementation of POS tagging and lemmatization.

Remove stop words: Stop words are commonly used words like 'I', 'she', 'the' or 'a', which have little meaning and can be therefore removed. NLTK provides a list of such words. However, there are some more words that should be added to this list. For example, words corresponding to numbers ('one', 'two', ...), single letters ('z', 'y', ...), Twitter terms ('twitter', 'hashtag', ...) or even time or date terms ('yesterday', 'month', 'january' ...) are missing in this list. Such words were added to the original list after research to apply to tweet texts.

Remove not necessary whitespaces: By removing many disturbing factors it can happen that unnecessary spaces remain. These can be found with *^\s|\s\s+|\s$* and removed accordingly.

Before lemmatization was done and stop words were removed, all letters were displayed as lowercase, because words were changed or searched for in these two steps and they were needed consistently. In the end, some rows of tweets were able to be deleted entirely, as for these texts filters removed all content, leaving only an empty tweets remained. Table 4.2 shows examples of how tweets have changed with all the filters.

In these examples one can see the results of the different steps. Hyperlinks, hashtags, user tags and conversion errors were removed from the raw tweets. Also various stop words are no longer included in the cleaned tweet. It can also be seen that lemmatization has provided good results and that meaningful words have still come out as output.

Raw Tweet	Cleaned Tweet
Contribute to the #GlobalGoals and save water by taking shorter showers. Every #ClimateAction counts. #ActNow and log yours today. https://t.co/TYrOUsk2JW	contribute save water take short shower every count log
Artificial intelligence could revolutionize sea ice warnings https://t.co/P8fejDbDwH via @CAGE_COE @EurekAlert #climatechange #AI	artificial intelligence could revolutionize sea ice warning
Completely agree &amp; I'm hoping people in power will stop rejecting scientific expertise full stop. #ClimateCrisis https://t.co/v3XEGO39l7	completely agree hop people power stop reject scientific expertise full stop

Table 4.2: Examples of cleaned tweets

4.2.2 Similarity Cleaning

In addition to cleaning up the texts, it also needs a look at the similarity of tweets. Twitter can be used for different purposes. Users can share their thoughts about all kinds of events, but Twitter is also sometimes used for tracking or marketing purposes. This is where bots come into play, for example, posting weather data at a certain time every day or bots posting flight data. Many companies use Twitter to promote stuff for their benefit. These can be advertisements for products or training courses. There are a number of other examples, but they all have one thing in common, they are disturbing factors in text analysis.

For this purpose, such tweets must be identified and sorted out, so that the amount of these tweets is kept at a minimum. One way of doing this is to compare the tweets with each other in order to identify similarities in tweets. A distance or difference of tweets can be calculated with the Levenshtein distance. Given two strings, the Levenshtein distance calculates the minimum number of inserting, deleting, and substituting operations to get from one string to the other[L+66]. The Levenshtein distance can be used for duplicate string detection and provides a suitable solution for the existing use case. Python provides a module[4] that can be used to calculate such Levenshtein distances. In addition to the distance, the ratio is also of interest, which is calculated from the ratio of the Levenshtein distance to the alignment length. The higher the value of the ratio, the higher the similarities of two strings.

The removal of similar tweets was implemented by first looking at each tweet to see which user this tweet is from, and then comparing all of this user's tweets with the currently selected tweet. Here, a ratio was calculated using the Levenshtein distance. Each tweet for which the ratio exceeded a threshold of 0,7 compared to the currently selected tweet was included in a list of references with similar tweets for the selected tweet. Thus, for

[4] https://pypi.org/project/Levenshtein/(Last accessed on 2022-08-29)

all tweets, all other tweets of a user could be found that have a high similarity. Since there may be one or more similarities, groups were defined based on the referenced tweets in the lists. This was further used to remove similarities, as within these groups it was calculated which tweet should continue to be preserved. This was found out by looking within the groups which tweet occurs most often in the list of similarities of the other tweets. If multiple tweets occur the same number of times, the longest tweet was chosen to persist because it contained the most information. All other tweets in this group were deleted from the dataset.

With this variant, similar tweets can be identified and removed. Since not all such tweets come from bots or companies, one tweet was always kept within the similarity group. Normal users can also send similar tweets, for example by making a small correction to a previous tweet and posting it again. If one removes all tweets that fall into such similarity groups, one would lose a certain amount of information content. This was to be addressed with the described approach.

4.2.3 Exploration of Data after the Text Preparation

After removing similar tweets as defined in the previous section and tweets that had no more words after a cleanup, the data amount can be evaluated again. In table 4.3 the top 16 European capitals listed based on the number of tweets after the text preparation. The column 'Difference' shows the difference from the evaluation of the dataset in table 4.1, which showed the results after city assignment. Additionally the column 'Percentage' shows the decrease of tweets in percent.

The table shows that there has been a significant reduction in the amount of data due to the text preparation of the tweets. The top 16 cities in this table are the same 16 as in the previous table. In absolute terms, of course, London has the most tweets less than before. But relatively, London is also in the top spot. London and Berlin have relatively the most tweets less, with over 15% fewer tweets than before. But also Stockholm and Athens show a high decrease with around 13%. In the case of London and Berlin, however, it was to be assumed due to the large amount of data that there are lots of tweets with similar content. In contrast, Brussels with the second most tweets has under 8% decrease of tweets, which is lower than the average amount of about 9%. Oslo with just under 3% and Copenhagen with mid 4% show the smallest decrease.

In the Data Understanding phase, the 5 users with the most tweets were looked at. There it could already be determined that the user with the most tweets named *chriscartw83* is partially doing data tracking with his account. For this user, the number of his tweets has decreased from 9.294 to 4.082, a decrease of 56%. This shows that similarity cleaning has had a significant effect.

Even after the Text Preparation step, a good balance between coastal and inland cities can be seen. Thus, eight cities by the sea and eight cities inland have a reasonable amount of data for further analysis. However, there is a slight imbalance in the distribution of cities within this ranking, as inland cities tend to be found in the upper part and seaside

City	Tweets	Difference	Percentage	By the Sea?
London	144.172	-28.118	-15,58%	no
Brussels	28.113	-2.773	-7,85%	no
Berlin	16.482	-3.154	-15,36%	no
Dublin	12.368	-957	-6,41%	yes
Paris	11.566	-978	-7,14%	no
Amsterdam	5.123	-566	-9,3%	yes
Rome	4.996	-559	-9,31%	no
Vienna	4.825	-529	-9,37%	no
Helsinki	4.316	-296	-5,5%	yes
Stockholm	4.018	-666	-13,46%	yes
Madrid	3.771	-459	-10,04%	no
Copenhagen	3.427	-180	-4,38%	yes
Oslo	2.254	-92	-2,93%	yes
Athens	2.244	-355	-12,65%	yes
Luxembourg	2.173	-186	-7,45%	no
Lisbon	1.515	-125	-6,6%	yes

Table 4.3: Top 16 amount of tweets per European capitals after text preparation

cities tend to be found in the lower part of the ranking. This can be balanced out in the analysis, though a balance in the number of cities is more significant.

4.3 Conclusion and selection of Dataset

In this phase, further steps were taken for a successful data analysis. The custom-created city dataset was linked to the collected tweets in order to determine the origin of the user's tweets. It was not possible to define general rules that would have applied to all cities, since many cities had certain characteristics that had to be taken into account. Nevertheless, a solution could be found to counteract the findings that had been discovered. Although this required a review of all the characteristics, it brought the desired results. The data exploration after the assignment of the cities showed that for cities with additional names in official language an increase of the assigned tweets could be detected, despite additional constraints from the findings. For cities without additional names in the official language, a decrease in the amount of data was observed with additional constraints.

Following this, the tweets themselves could be prepared for analysis. This was done using Python implementations. The texts were analyzed and accordingly steps were applied to remove disturbing factors from the tweets. The disturbing factors, such as hyperlinks, were removed using regex. Using lemmatization, the words were brought to a common base, and subsequently stop words were removed from them. Thus, a cleaned version of the tweets could be obtained. Furthermore, similar tweets could be explored and removed from the dataset. The results of these steps showed a decrease in the dataset of 2,93-15,58% per city with an average of 8,96%.

Overall, after the assignment to cities and the cleaning process, 259.737 tweets could be assigned to a European capitals, 251.363 to the cities in the top 16 alone. Since there are an equal number of coastal and inland cities in the top 16 (eight to eight), it makes sense to select the top 16 for the further analysis process.

CHAPTER 5

Modeling

In this phase the appropriate methods and models are selected, which with their results are going to lead to answering the research questions. Thereby the two topics polarizing words and sentiment analysis are discussed. The development of the methods and models provides data for the Evaluation phase, in which the data is analyzed.

5.1 Polarizing Words

Polarizing words can be identified and displayed in different ways. With the help of word clouds, information about the frequencies of words in texts can be displayed visually [HLLE14]. This helps to get a quick overview of the frequency and importance of words in a context. However, word clouds are used to represent a large number of terms, not all of which can be considered polarizing terms, since they occur rather rarely. The larger words are represented in word clouds, the more frequently they occur. Since the goal is not only a representation of the polarizing words, but also a comparison between different European capitals, it requires additional methods. Therefore TF-IDF (term frequency-inverse document frequency) is chosen, which can determine relative frequencies of words in texts to compare documents with each other [HMB17, MRHC21]. Here it is especially important to compare the relative number of words in relation to the full text. One can see in the amount of tweets that some cities have significantly more tweets than others. An absolute calculation would lead to large differences in the comparison if there were large differences in the amount of data. However, absolute comparisons are still made in the top words of the respective cities with each other to find out if the top words in the cities are the same and to present them.

5.1.1 WordClouds

The representation of the data using word clouds is done with the Python module wordcloud[1]. By default, the parameter *collocations* is set to *true* when using this module. This parameter specifies whether collocations (bigrams) of two words should be used above a certain occurrence. The words climate and change would be such a collocation, because these words often occur in this order, but not always. Since this is not always the case, this function is deactivated, because otherwise 'climate', 'change' and 'climate change' would occur in the word clouds and thus 'climate' and 'change' would look as if these two words do not occur as often as they actually do.

5.1.2 Relative Comparison of Texts

In order to be able to compare the European capitals with each other, first the tweets per city were summarized into texts, so that each city has a text with all tweets of this city. These texts can also be considered as documents to be compared. TF-IDF (term frequency-inverse document frequency) is a numerical statistic [QA18] and is often used in the field of text classification to compare relative frequencies in texts or documents using inverse proportion of terms over the all texts or documents to be compared [HMB17]. It is mostly used to find the relevance of certain words in documents [QA18], but is also used to check documents for similarity by vectorization [MRHC21].

TF-IDF consists of two term combinations: term frequency (TF) and inverse document frequency (IDF). TF measures the number of occurrences of a word in a document relative to the document size, giving a relative occurrence frequency of a word. IDF measures the importance of a word over all documents using inverse proportion of the words. As a result, the product of the two measurements is TF-IDF. [HMB17, QA18, MRHC21]

Since the tweets were already created as texts by city at the beginning, they can be compared using TF-IDF to determine similarities in the texts. These similarity values can be calculated with the use of the sklearn module in Python[2] which provides such a class of TF-IDF calculations. Within this module the *TfidfVectorizer()* is used to turn the words into vectors. As data preparation was already made, most of the parameters for preparation purposes can be skipped. However, there are two parameters that need to be looked at for the purpose of this analysis. They are *use_idf* and *smooth_idf*. IDF is used to give more weight to words with low occurrences, so that they do not go down but gain importance [HMB17]. *use_idf* activates the weighting of words and would have IDF=1 when deactivated and *smooth_idf* adds one to the document frequencies so that zero divisions can be prevented. Keeping this active is mostly for the purpose of searching for certain words and their relevance. However, when comparing the tweets of the respective cities, this is not very relevant and therefore these parameters can be deactivated, as they are activated by default. A simple example shows the effect of deactivation and the results of the comparisons. When comparing the texts 'water ocean' and 'water sea'

[1] `https://pypi.org/project/wordcloud/`(Last accessed on 2022-10-31)
[2] `https://scikit-learn.org/stable/`(Last accessed on 2022-10-31)

with both parameters activated, a similarity value of 0.34 is obtained. Whereas with deactivation the similarity value 0.5 results, meaning a 50% difference of the two texts, which actually exists to the extent. This can now be used to compare tweets among each other for similarity.

5.2 Sentiment

Sentiment analysis or sometimes opinion mining is a method in the field of Natural Language Processing (NLP)[HG14, MSM22]. It is used to identify and categorize people's sentiment, opinions, emotions and attitudes in texts [HG14, BKJ19] to find out whether the texts have positive, neutral or negative sentiment [BKJ19, DKL19]. In the field of sentiment analysis there are two possible approaches to choose from, namely the lexicon based approach and the machine learning approach. The lexicon-based approach uses models that have been trained and that already rate words or phrases based on sentiment lexicons and in contrast, the machine learning approach uses statistical methods and training data to create models [BKJ19, MSM22].

Two well-known examples of lexicon based approaches are TextBlob and Valence Aware Dictionary and sEntiment Reasoner (VADER). VADER was developed by Hutto and Gilbert [HG14] and is a sentiment model for social media texts, especially microblogging like Twitter is. Textblob, on the other hand, is for a more general context [EKSA22]. Comparisons of these two sentiment analyses also show that VADER is better suited for Twitter data than TextBlob [EKSA22, BKJ19].

Since VADER is already a good model for sentiment analyses in the field of Twitter, this model can also be applied to the present data, especially since Effrosynidis et al [EKSA22] made the assessments with Twitter data in the field of climate change. The use of VADER has several advantages [HG14, BKJ19]:

- The calculation of the results is quick and computationally economical
- It does not require training of the data, as it already performs well in various domains.
- It is specialized on social media texts, especially microblogging as Twitter is

Especially the last point is characterized by the identification of human heuristics in texts, as it increases the magnitude of the sentiment intensity because of punctuation or ALL-CAPS phrases among others [HG14]. This leads to the conclusion that in a sentiment analysis with VADER it is better to use the original tweets rather than preprocessed tweets. In the Data Preparation phase, not only the tweet texts were cleaned, but also similar tweets were filtered out. This is an important part that should be considered for the sentiment analysis. Therefore, from the file with the preprocessed data, in which similar tweets were filtered out, the original tweets of the respective preprocessed tweets were taken.

The result of the VADER sentiment analysis is a *compound score*, which indicates how negative or positive a tweet is in a range from -1 (negative) to +1 (positive). A typical threshold value for the classification is *compound score <= -0,05* as negative, *compound score > -0,05 and < 0,05* as neutral and *compound score >= 0,05* as positive [BKJ19].

CHAPTER 6

Evaluation

This phase presents the results of the analyses for the research questions and subsequently answers the research questions. This phase is divided into two sections, *Polarizing Words* and *Sentiment*, in which the respective research questions are addressed.

6.1 Polarizing Words

In this section, the polarizing words of the tweets are observed and comparisons between cities are made. In particular, the comparison is made between the groups of coastal cities and inland cities. With these analyses and evaluations, the following research question will be answered: ***Which terms in the context of climate change polarize on Twitter in European cities on the coast and in European inland cities and do these groups differ?***

Therefore it is necessary to find out which words polarize in general and in the European capitals and whether they are similar with the same group of cities. With groups inland and coastal cities are meant. To find out and display the polarizing words, word clouds are used among others. In addition, the top words of the respective cities are also considered and compared with each other to find out which words polarize in several cities and which words polarize possibly only in certain cities. For the comparison of the cities, however, it is important not only to make absolute comparisons, but also relative ones. As already seen in analyses in other phases, the number of tweets per city differs. For example, London has significantly more tweets (144.172) than Lisbon (1.515). Therefore, the frequencies are considered in relation to the number of tweets within the cities, since 1.000 occurrences in one city could be equivalent to 100 occurrences in other cities. These comparisons are made using TF-IDF.

First let's look at the word cloud over all texts of the top 16 cities, which can be seen in figure 6.1. The ten most common words seen in the word cloud are 'climate', 'change',

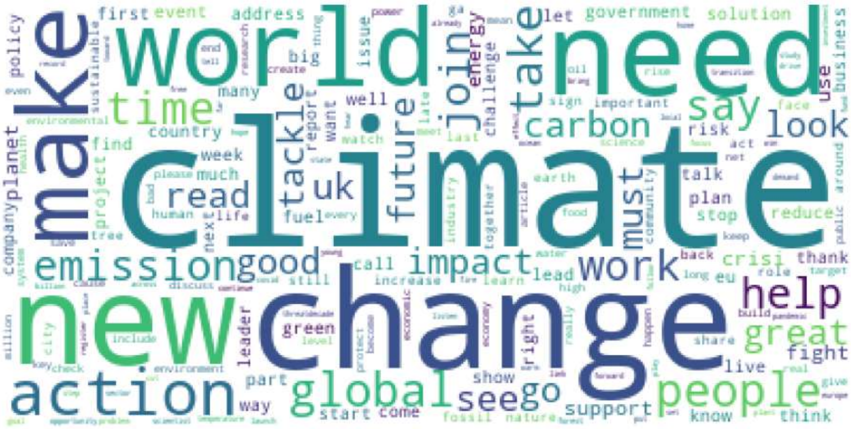


Figure 6.1: Wordcloud of texts including all cities

’need’, ’new’, ’world’, ’make’, ’action’, ’global’, ’take’, ’people’. It can be seen that the word ’climate’ is quite dominating. This can also be seen in terms of numbers, as ’climate’ occurs 52.393 times, followed by ’change’ with 29.935 and ’need’ with 22.030. It is also interesting to see what kind of words they are and what they are about. Of course, in terms of meaning, the words relate to climate change, even in a further sense. Since there are not only nouns in the top 10 words, but also action words that can be related to climate change in a way to tackle problems, such as make or take in combination with action or change. From this it can be seen that climate change is progressing from the point of view of humanity and action is needed to counteract it. Then climate change is a very defining topic and concerns people all over the world.

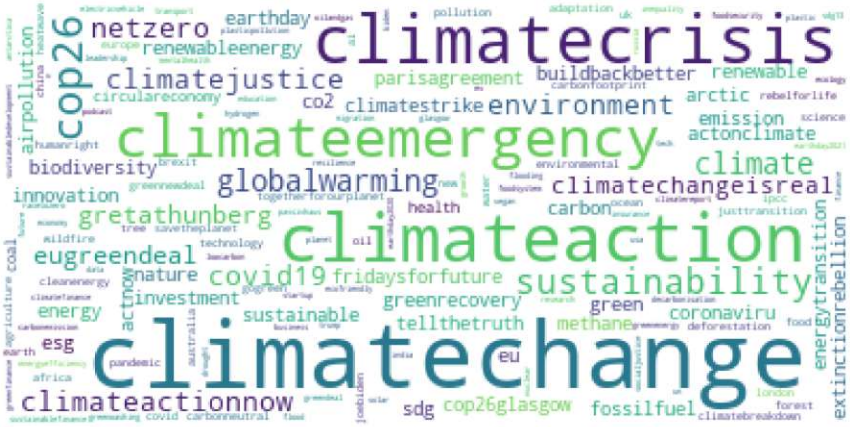


Figure 6.2: Wordcloud of hashtags including all cities

In the Data Understanding phase the existing hashtags were already analyzed and displayed as a list in table 3.3. The word cloud in figure 6.2 serves as an additional graphical representation of this. As already shown in table 3.3, the hashtags 'climatechange', 'climatecrisis' ad 'climateaction' are the ones with the most occurrences. In addition some more hashtags can be seen that were not included in the table. However, it should be noted that the number of occurrences is not entirely based on the table, since in between also data cleansing was done in the Data Preparation phase and thus one or the other tweet and so also some hashtags were filtered out.

These two word clouds are over the whole remaining dataset and thus have a slight bias due to absolute occurrences in the respective cities, as in some cities there are more tweets than in others. So in a next step it is interesting to see which words just occur most frequently in the cities. For this purpose, the top ten words from each of the 16 cities are taken and summed up over all cities. Since there are 16 cities, there can be a maximum of 16 occurrences of a word. In figure 6.3 these words are represented by a bar plot.

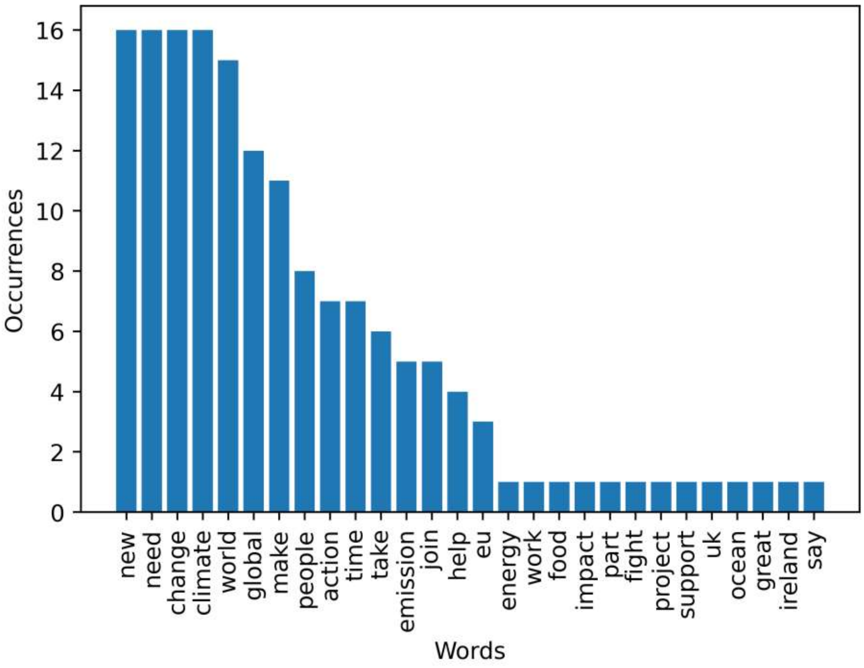


Figure 6.3: Barplot with the summed up top 10 words of all cities

In this bar plot, one can see that the ten most common words not only occur most frequently in the word clouds, but also have the most frequent occurrences in the respective top 10 words over all cities with just one exception. The word 'time' has seven occurrences and is therefore more frequent present in the top 10 words of all cities than 'take' with six occurrences. The words 'climate', 'change', 'need' and 'new' even appear in all 16 cities in their top 10 words. Furthermore, it can be seen that some words occur rarely or even only once. Comparing the top 10 words of all cities with each of the other top 10 words of the other cities, the average equality value is 6,83 words. However, to have more words and thus more expressiveness, the same is done again with the top 20 words of the respective cities, which can be seen in figure 6.4.

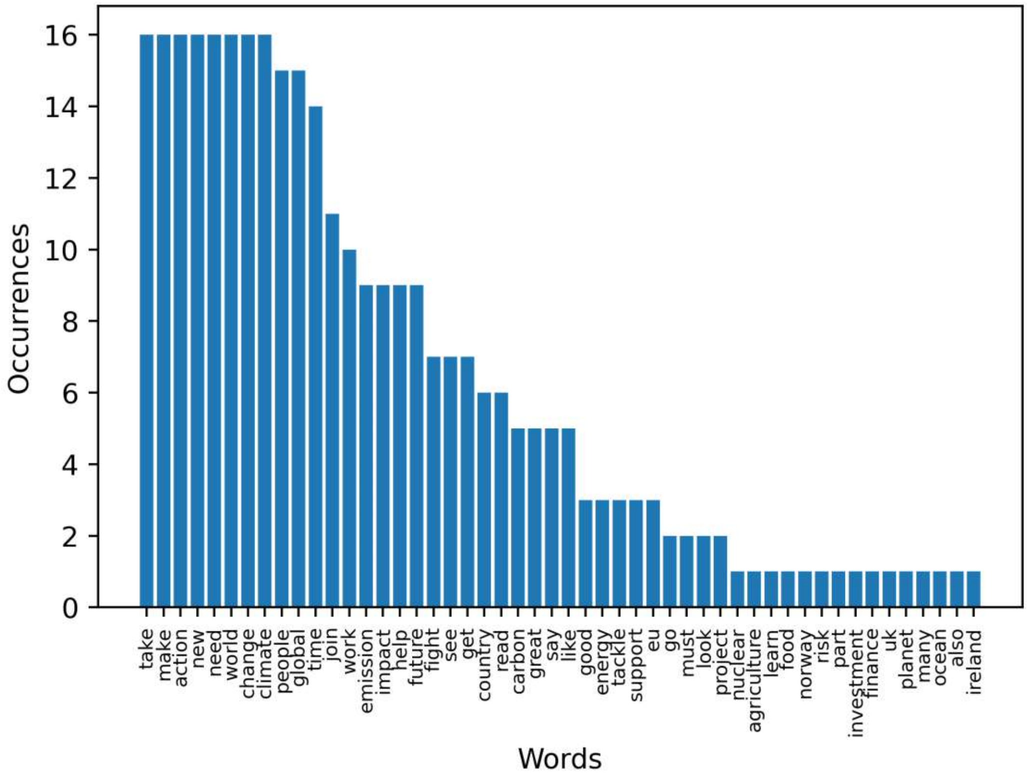
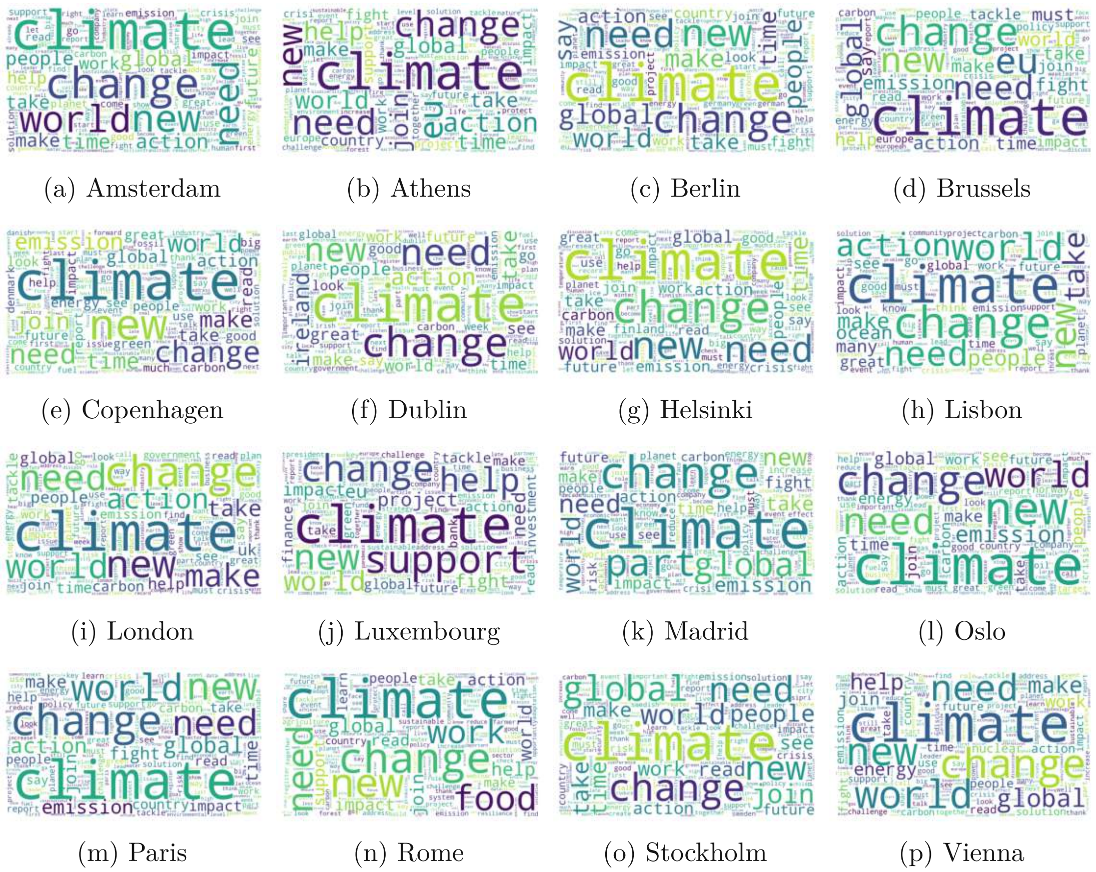


Figure 6.4: Barplot with the summed up top 20 words of all cities

Looking at the bar plot with the top 20 words summed up, it can be seen that this time the ten most common words from the word cloud are without exception also the ten most common words in the bar plot. Further, only the words 'global' and 'people' do not appear in one city each. Besides that all other words of the ten most common words of the word cloud are represented in the top 20 words of each city. The average equality of words compared to other cities is 13,81. Going through the words, one can see three words that can be clearly assigned to cities, namely the countries in which the cities are located ('ireland' to Dublin, 'uk' to London and 'norway' to Oslo). For the other words this is not so easy anymore, so let's have a look at the word clouds of the individual cities in figure 6.5.

In general, it can be seen that 'climate' and 'change' dominate in each city and that there are high similarities between the word clouds. This fits with the evaluation in the previous bar plots (figure 6.3 and figure 6.4), since the most frequent words are roughly the same in all cities or occur in the top 10 respectively top 20 words of the cities. Occasionally, there are words that occur frequently only in certain cities, which are shown with a low occurrence in the bar plots. In looking at the words which occur



(a) Amsterdam (b) Athens (c) Berlin (d) Brussels

(e) Copenhagen (f) Dublin (g) Helsinki (h) Lisbon

(i) London (j) Luxembourg (k) Madrid (l) Oslo

(m) Paris (n) Rome (o) Stockholm (p) Vienna

Figure 6.5: Word clouds of texts per European capital

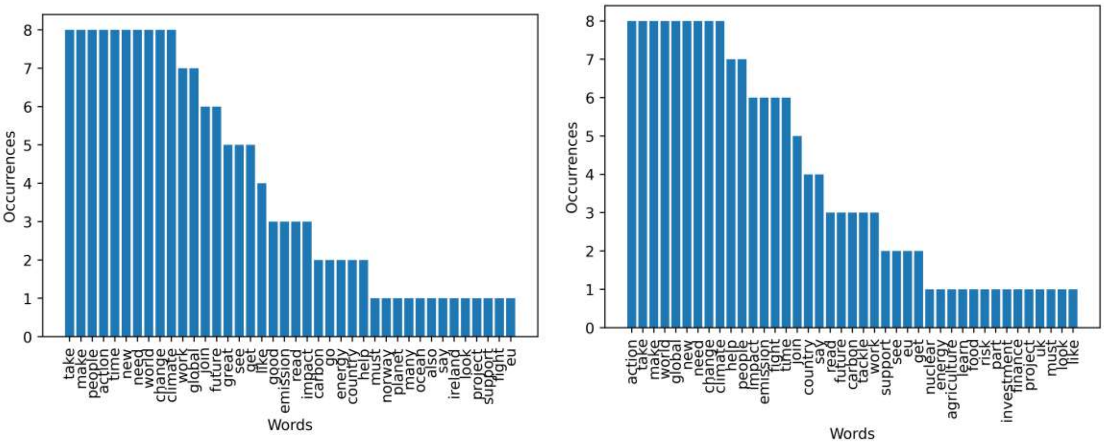
only once, among others, the above mentioned countries can be found in the word clouds at the cities of these countries. But also other interesting words occur only at a certain city frequently.

In the bar plot of the top 20 words in figure 6.4, the words ’ocean’ (Lisbon), ’planet’ (Lisbon), ’finance’ (Luxembourg), ’investment’ (Luxembourg), ’food’ (Rome), ’agriculture’ (Rome) and ’nuclear’ (Vienna) only appear once in the top 20 words. In Lisbon, there are many tweets that deal with the planet, especially the ocean, which seems appropriate, since Lisbon is located by the sea. In Luxembourg, the topic of climate change seems to revolve around the world of finance, as the words ’finance’ and ’investment’ appear frequently because of the relatively high number of banks in this country. In Rome some tweets are about ’agriculture’ and ’food’, which highlight the importance of agriculture and quality of food in Italy. Although Austria has no nuclear power, the word ’nuclear’ occurs frequently because the International Atomic Energy Agency (IAEA) has its headquarters in Vienna and some tweets originate from there.

Since until now not so much attention has been paid to the groups with coastal and inland cities, a first overview for this shall be provided by a comparison. In figure 6.6 the bar plot from figure 6.4 is split into coastal and inland cities. The grouping is as follows:

Coastal Cities Amsterdam, Athens, Copenhagen, Dublin, Helsinki, Lisbon, Oslo and Stockholm

Inland Cities Berlin, Brussels, London, Luxembourg, Madrid, Paris, Rome and Vienna



(a) Coast

(b) Inland

Figure 6.6: Comparison of words in coastal and in inland cities

The words already considered can be found in the corresponding groups. The words from Lisbon in (a) Coast and the words from Luxembourg, Rome and Vienna in (b) Inland. If one compares other words that appear in the top 20 words of the respective groups, one gets, for example, for 'eu' a result of 1:2 (Coast:Inland). Also for 'energy' (2:1) and 'carbon' (2:3) the results are rather close, so that one cannot conclude much from it. On the other hand, some other words show larger differences between the groups. For example, the words 'future' (6:3) and 'great' (5:0) tend to occur more often in coastal cities, whereas the words 'fight' (1:6), 'help' (2:7), 'impact' (3:6) and 'emission' (3:6) occur more often in inland cities. It is clearly noticeable that e.g. the words 'carbon' and emission', which have a similar background, also show a similar tendency. All in all, one can already see first similarities within the groups that differ from the other groups.

To further check for similarity, approaches are needed to compare the data sets. This following approach is used for the comparison of the top 20 words and in later stages similarly for the relative comparisons of the data sets. For this purpose a 16x16 matrix is built with all cities and in the matrix it is compared how many words are equal when comparing the respective two cities in the matrix. This results in a 16x16 matrix where each city has a comparative value to all other cities. The cities are then divided into their two groups and looked at each city to see if there are more similarities for that city to cities in the same group or to cities in the other group. Two approaches were chosen to evaluate the similarity:

1. Comparative Value: An average is calculated for the two groups using the comparative values from one city to all other cities in the respective group, i.e. an average for the cities in the same group and an average for the cities in the other group. The higher the value, the more identical words there are and the more similar the city is with this group.

2. Ranking: The comparative values are sorted in descending order and a ranking from 1 (highest similarity) to 15 (lowest similarity) is made. If two or more comparative values are the same, all of them receive the same rank and the next value is one higher. The rank is averaged for each group and the lower the average rank, the more similar the city is to that group. This should help not to let outliers fall too heavily into the weight.

This therefore gives two similarity tests for each city. Since there are eight cities in each group and there are two similarity tests for each city, there are 16 scores for groups showing which group has more similarity.

City	By the Sea?	Comparative Value	Ranking
Lisbon	yes	16	1
Amsterdam	yes	15	2
Helsinki	yes	15	2
Berlin	no	14	3
Copenhagen	yes	14	3
Oslo	yes	14	3
Stockholm	yes	14	3
London	no	13	4
Madrid	no	12	5
Vienna	no	12	5
Athens	yes	11	6
Brussels	no	11	6
Paris	no	11	6
Rome	no	10	7
Luxembourg	no	9	8
Mean values coast		14,14286	2,85714
Mean values inland		11,5	5,5

Table 6.1: Evaluation of the comparisons with Dublin

To make this more concrete in the case of the top 20 words, such a matrix was created. The average equality value is 13,81 words. When comparing two cities, a maximum equality of 17 words and a minimum equality of nine words was obtained. Using the example of Dublin in table 6.1, it can be seen how the comparison was made. First, there is the *Comparative Value*, which maps the first approach and lists the comparative values. On the other hand there is the *Ranking*, which gives a rank to the sorted comparative values. Finally, an average is taken over both columns depending on the membership of the group. In this example, we can see that Dublin shows a clearer similarity both in the comparative value and in the ranking in its group, i.e. the coastal cities, as the comparative values is higher and the ranking is lower.

	Comparative Value		**Ranking**		**Score**	
City	**Coast**	**Inland**	**Coast**	**Inland**	**Coast**	**Inland**
Amsterdam	**15,42857**	14,625	**2,57143**	3,375	2	0
Athens	13	**15,125**	3,71429	**1,875**	0	2
Copenhagen	**14,57143**	13,25	**2,42857**	3,625	2	0
Dublin	**14,14286**	11,5	**2,85714**	5,5	2	0
Helsinki	**14,85714**	13,25	**2,14286**	3,625	2	0
Lisbon	**13,85714**	11,625	**3,14286**	5,375	2	0
Oslo	**14,85714**	13,75	**2,14286**	3,125	2	0
Stockholm	**15**	14	**2,14286**	3	2	0
Overall	**14,46429**	13,390625	**2,64286**	3,6875	14	2

Table 6.2: Evaluation of similarity of top 20 words in coastal cities

	Comparative Value		**Ranking**		**Score**	
City	**Coast**	**Inland**	**Coast**	**Inland**	**Coast**	**Inland**
Berlin	**14,25**	14	**2,75**	2,85714	2	0
Brussels	13,125	**15**	4,625	**3**	0	2
London	13,875	**14**	3,125	**3**	0	2
Luxembourg	11,125	**12,42857**	4,875	**3,57143**	0	2
Madrid	13,625	**14,28571**	3,375	**2,71429**	0	2
Paris	13,75	**15,57143**	4	**2,42857**	0	2
Rome	12,875	**13,14286**	3,875	**3,85714**	0	2
Vienna	**14,5**	14,42857	**2,25**	2,42857	2	0
Overall	13,39063	**14,10714**	3,60938	**2,98214**	4	12

Table 6.3: Evaluation of similarity of top 20 words in inland cities

In the next step, this is done for all cities. In the tables 6.2 and 6.3, the evaluation of similarity is shown separately for the coastal cities and the inland cities. These tables show the average values of the two approaches and, in addition, a score indicating which group the results of each city and overall were in favor of.

The results show that there are more similarities within their own groups than with cities of the other group. For coastal cities, the score is 14:2 in favor of cities in their own group, which indicates that all but one of the cities are more similar to cities in the same group than to cities in the other group. The exception in this case is Athens, where both approaches indicate that there are more similarities with cities of the other group. For inland cities, it is similar, but not quite as clear as for coastal cities. Here the score is 12:4 in favor of cities in their own group, with two out of eight cities as exceptions. Berlin and Vienna are more similar to the cities in the other group. Both groups also show clear tendencies towards their own group on the overall averages for comparative values and rankings.

Going a step further, the occurrences are also compared in relation to the number of tweets in the cities. A comparison of all words does not necessarily correspond to the question about polarizing words, since this would also include many words that only occur once. Therefore, such words must be removed. However, removing words that occur less than a certain number is not a solution. This would affect cities with fewer tweets significantly more, since there are not so many words there and proportionally significantly more words would be removed. For example, if all the words that occur less than ten times were removed, Lisbon would have about 90% of all words removed, whereas London would have only about 80% removed. Since we want to compare proportionally, a better approach is to remove the same percentage of words for all cities.

In order to get a feeling for how many words occur how often, lets look at different percentiles. For this purpose, all words are counted by occurrence and sorted in ascending order. If one looks at the frequency of the words at the 50th percentile, i.e. the median, in this sorted list, it results that at the median the frequency of the words of the respective cities is either once or twice and on average over the 16 cities at 1,75. This is a too small number to cut off the words at this point. Since we have already seen that in Lisbon at the 90th percentile this value was at ten occurrences, this is considered next. At the 90th percentile, truncation would occur at ten (Lisbon) to 36 (London) occurrences, averaging 17,81. However, this would still leave a very high number of words left over, as it ranges from 403 to 4.273 (on average 1083,69) words to be left over when truncated at the 90th percentile. When approaching a corresponding value, at the 97,5th percentile, each city would still have over 100 words left over, but only barely. This way, for the cities 109 to 1.065 (on average 270,19) different words would remain, removing words with less than 26 to 311 (on average 83,19) occurrences. Thus, not too few, but also not too many words would have to be removed.

Therefore, let's take a closer look at the results of the words that are above the 97,5th percentile to make comparisons between the cities. For this purpose, using TF-IDF, the remaining words are compared for similarity to all other cities, taking into account the frequency in relation to the total number of remaining words per city. The two approaches (comparative value and ranking) are again used for the evaluation.

Using TF-IDF, comparing two cities results in a score between 0 and 1, where 0 means no similarity and 1 means exact similarity. In figure 6.7 the box plots of the cities can be seen and how similar they are in comparison to the other cities. Regardless of

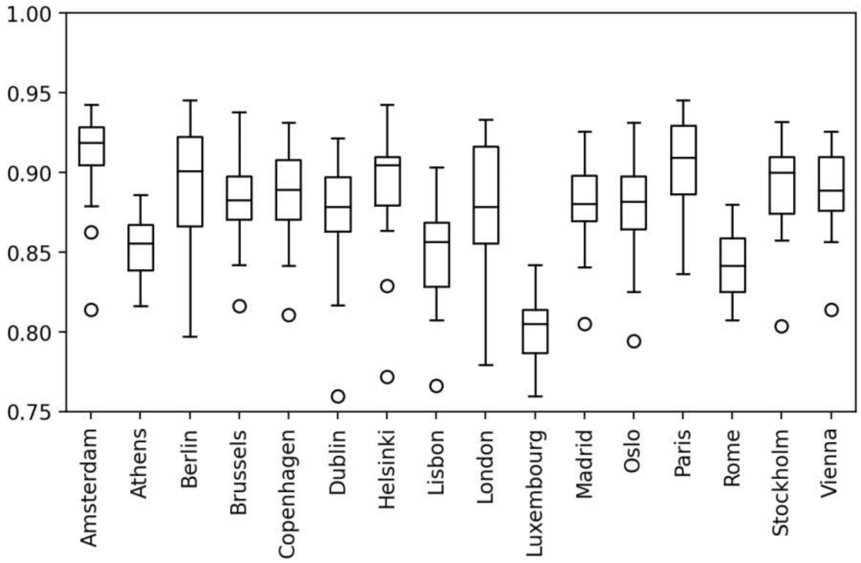


Figure 6.7: Similarity per city compared to all other cities

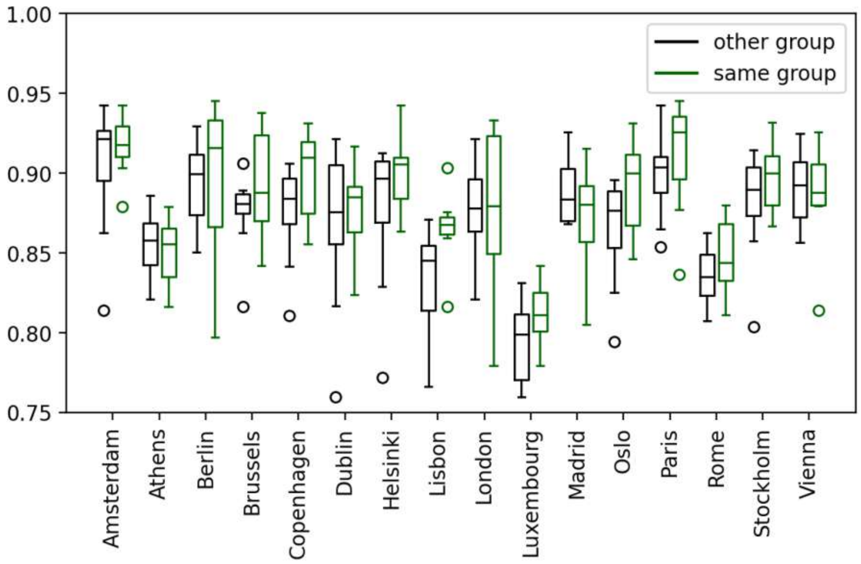


Figure 6.8: Similarity per city compared to cities of the same and of the other group

location, whether the city is on the coast or inland, one can see that the similarity is very high in most cases. Especially Amsterdam and Paris show a high similarity to all other cities. Rather less similarity with the other cities have Athens, Lisbon, Rome and in particular Luxembourg. In the case of Luxembourg, it can be seen that there are only low similarities with all the other cities. Also, the extreme downward outliers that can be seen in the plot is Luxembourg in most cases. This is because Luxembourg generally has rather different words, which can already be seen in table 6.3 at the comparative value. There, the average values for coastal and inland cities of Luxembourg are significantly lower than all others, which means that the top 20 words of Luxembourg are more likely to have other words than the other cities. This is then also reflected in the words above the 97,5th percentile.

In figure 6.8 a comparison per city can be seen, how similar this city is to the cities of the two groups. The tables 6.4 and 6.5 then show the associated data.

	Comparative Value		**Ranking**		**Score**	
City	**Coast**	**Inland**	**Coast**	**Inland**	**Coast**	**Inland**
Amsterdam	**0,91706**	0,90305	**7,71429**	8,25	2	0
Athens	0,8503	**0,85511**	8,71429	**7,375**	0	2
Copenhagen	**0,898**	0,87488	**6,28571**	9,5	2	0
Dublin	**0,87656**	0,86685	**8**	**8**	1,5	0,5
Helsinki	**0,9**	0,876	**7,28571**	8,625	2	0
Lisbon	**0,86535**	0,83279	**5,28571**	10,375	2	0
Oslo	**0,89087**	0,8636	**6,28571**	9,5	2	0
Stockholm	**0,89731**	0,88081	**7,14286**	8,75	2	0
Overall	**0,88693**	0,86914	**7,08929**	8,79688	13,5	2,5

Table 6.4: Evaluation of similarity of words above the 97,5th percentile in coastal cities

	Comparative Value		**Ranking**		**Score**	
City	**Coast**	**Inland**	**Coast**	**Inland**	**Coast**	**Inland**
Berlin	0,89215	**0,89412**	8,75	**7,14286**	0	2
Brussels	0,87516	**0,8937**	9	**6,85714**	0	2
London	**0,87725**	0,87697	8,375	**7,57143**	1	1
Luxembourg	0,79414	**0,81216**	9,5	**6,28571**	0	2
Madrid	**0,88865**	0,87132	**7,5**	8,57143	2	0
Paris	0,89851	**0,91027**	9,25	**6,57143**	0	2
Rome	0,83589	**0,84833**	9	**6,85714**	0	2
Vienna	**0,89133**	0,88553	**8**	**8**	1,5	0,5
Overall	0,86913	**0,87405**	8,67188	**7,23214**	4,5	11,5

Table 6.5: Evaluation of similarity of words above the 97,5th percentile in inland cities

In the box plot it can be seen that in most cases there is a tendency towards more similarity to cities in their own group. In the two tables (6.4 and 6.5) the two approaches are shown for comparison, which were already used for the top 20 words. And they also show a very similar picture in the results. As most of the cities have more similarity to cities of their own groups. In coastal cities, the total score is 13,5:2,5 in favor of cities of their own group and in inland cities it is 11,5:4,5. The comma values come from the fact that in both coastal and inland cities there is one city where the ranking value is the same, namely Dublin and Vienna. It is also the case for these two cities that they show more similarities in comparative value to the cities in the other group. Vienna was also more similar to cities of the other group in the top 20 words comparison. In addition, the cities of Athens and Madrid also show more similarities to the other group in both approaches. In one case, London, the two approaches produce different results. While the comparative value suggests that London is more similar to cities in the other group, the ranking shows the opposite. This is due to the fact that Luxembourg has few similarities to other cities and when comparing London to Luxembourg, the comparison value clearly falls away from the other cities. Thus, the value in the evaluation for London is lowered for cities in the same group. For the remaining cities, there is more similarity to cities in the same group. Both groups show overall in the values that there is more similarity to cities in their own group. For the inland cities, the comparative value is closer together, but this is also due to the less similarities of Luxembourg to all the other cities, including those in its own group. For the ranking value, the results are just as clear for both.

In addition to the comparison of words above the 97,5th percentile, the same is done for nouns above the 97,5th percentile. Using the NLTK module[1], the nouns in the texts were identified and subsequently truncated at 97,5th percentile. Table 6.6 summarizes the overall results per group. It is interesting to see that for coastal cities, all cities show more similarity to cities in their own group for both approaches. For inland cities, the result is similar to that with all words above the 97,5th percentile. It are also the cities of London (1:1), Madrid (2:0) and Vienna (2:0) that show more similarities to cities of the other group. Luxembourg also plays a major role in the comparative value, since it also has a larger deviation for nouns than all other cities. The difference in comparative value for inland cities is thus very small.

	Comparative Value		**Ranking**		**Score**	
Group	**Coast**	**Inland**	**Coast**	**Inland**	**Coast**	**Inland**
Coast	**0,89057**	0,87	**6,82143**	9,03125	16	0
Inland	0,86986	**0,87191**	8,53125	**7,41071**	5	11

Table 6.6: Evaluation of similarity of nouns above the 97,5th percentile

In conclusion, there are some words related to climate change that characterize the European capitals. It could be seen that climate and change dominate in all word clouds of the cities and there are high similarities between the word clouds of the cities. This can also be seen in the top 20 words of the respective cities. There, the most common words (including 'climate', 'change', 'need', 'new', 'world') are represented in the top 20 words in all cities. When looking at words that do not occur in every city, first tendencies for the division into the two groups could be identified. So the words 'future' and 'great' could be found more in coastal cities and the words 'fight', 'help', 'impact' and 'emission'

[1] https://www.nltk.org/(Last accessed on 2022-11-21)

more in inland cities. To make a more precise comparison of the two groups, cities were compared and evaluated based on their top 20 words, all words above the 97,5th percentile, and all nouns above the 97,5th percentile. In table 6.7 one can see summarized the three methods with their score within the same group and to the other group.

	Score	
Method	**Same Group**	**Other Group**
Top 20 words	26	6
Words above the 97,5th percentile	25	7
Nouns above the 97,5th percentile	27	5

Table 6.7: Scores of the three methods for comparing similarity

It can be seen that very similar results could be obtained with all three methods, because similarities were detected within the groups. In general, the similarities in the comparison was very high, but it can be said that the two groups do differ. So the European capitals show significant more similarities to cities in the same group than to cities in the other group.

6.2 Sentiment

This section looks at the sentiment of tweets to answer the following research questions:

- ***Which three European cities have the most negative mood about climate change based on using the sentiment of tweets compared to other European cities and where are these cities located at?***
- ***Are there similarities in the context of climate change within European cities on the coast and within European inland cities based on using the sentiment of tweets and do these groups differ?***

For this purpose, comparisons are made between the cities and then between the two groups, inland and coastal. Sentiment analyses are made to determine the mood of the cities. With the help of the sentiment analysis it can be determined how the mood is currently in Europe and whether there are differences between inland cities and coastal cities.

Therefore, the compound score is calculated for each tweet, which subsequently allows the classification of tweets into positive, negative or neutral sentiment. This makes it possible to evaluate the tweets on the basis of a score and on the other hand on the basis of the classification of the tweets.

To get a first overview of the sentiment situation, box plots of all cities were created and compared, see figure 6.9. The values used for the evaluation are the compound scores of the tweets of the respective cities.

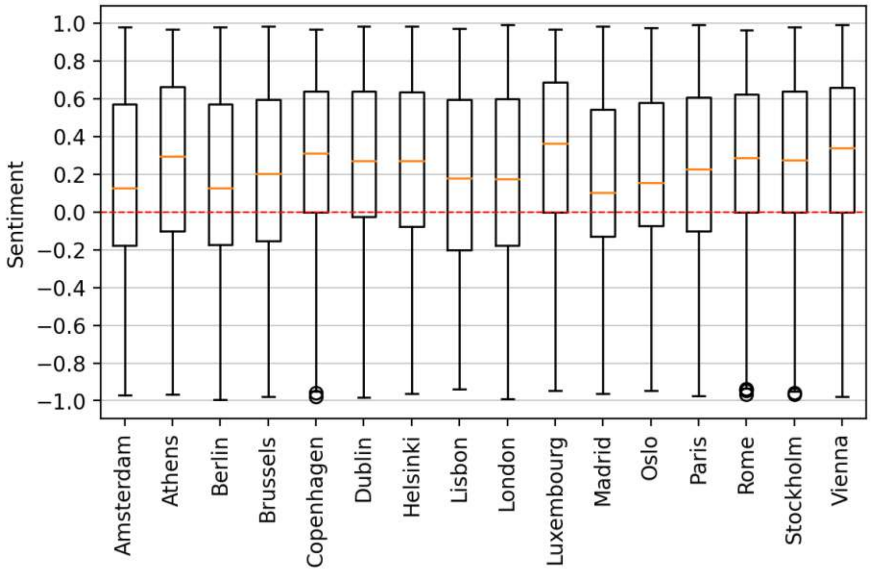


Figure 6.9: Box plots of compound scores per tweets of the cities

In figure 6.9 it can be seen that the compound score of the tweets for each city results in a positive sentiment. The median scores all hover around 0,2, and the average across all cities of 0,189 is also in this region. The 25th percentile for all cities is 0 or just below. The range of scores of the tweets for all cities ranges from very positive (1) to very negative (-1). However, no tweet reaches the maximum or minimum value of 1 or -1.

The differences between the cities is only slight, but some differences can be seen. The average values per city are between 0,14 and 0,26, with Amsterdam having the lowest value and Luxembourg the highest. In the previous evaluation of the polarizing words above the 97,5th percentile in figure 6.7 and the tables 6.4 and 6.5, it can be seen that exactly these two cities also have the highest and the lowest values of the comparative value. Luxembourg has the lowest similarity to the other cities and the most positive sentiment of all cities, whereas Amsterdam has the highest similarity to the other cities and the most negative sentiment of all cities.

Looking at the distribution of the classification of tweets across the entire dataset, the results from figure 6.9 are reflected in figure 6.10. In the pie chart, the majority of tweets (54,01%) are positive. Only 18,73% neutral and 27,26% negative tweets could be classified. Thus, more than half of the tweets have a compound score greater than 0,05 (positive) and just over a quarter have a compound score less than -0,05 (negative). The remaining tweets are neutral and fall between these two values.

In addition to the classification of the tweets, it is also interesting to look at the development over time. Figure 6.11 shows a time chart of how the compound score has developed over the months in 2020 and 2021.

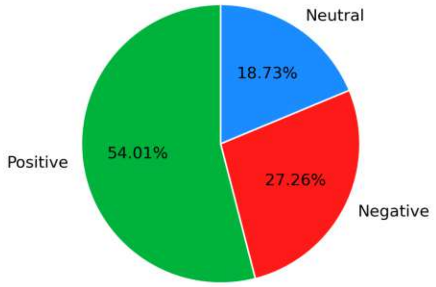


Figure 6.10: Pie chart with the classification of tweets for all cities

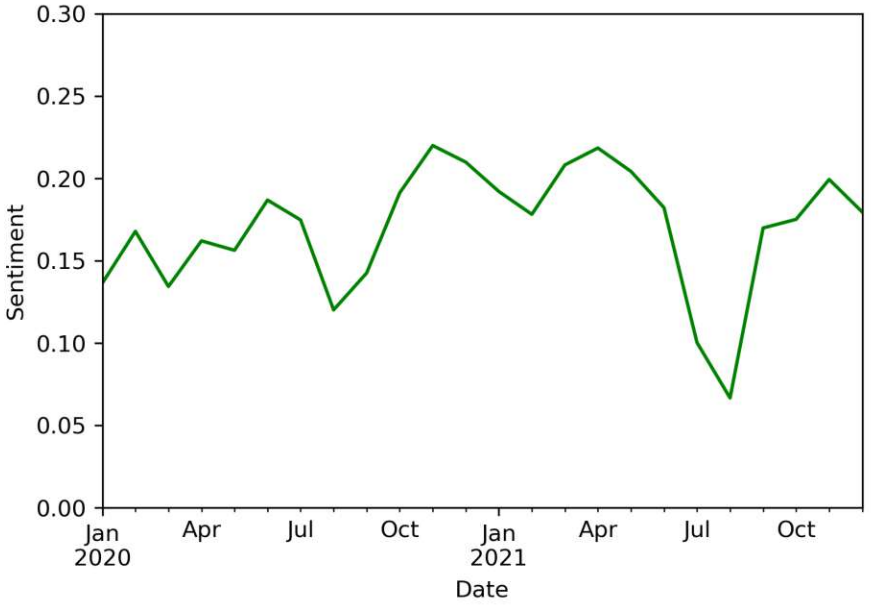


Figure 6.11: Average compound scores in the months of 2020 and 2021 from all cities

In these two years, neither a positive nor a negative trend is noticeable. There are fluctuations, which are particularly negative in the summer months. For both years, August is the month with the most negative sentiment, which may be due to heat waves.

Coming back to the values of the individual cities, let's take a closer look at the compound scores of the cities. Table 6.8 shows a ranking of the cities from the most negative compound score to the most positive compound score. The table shows in which country the cities are located and whether the cities are located by the sea. Also, the table shows the average compound score and the average ranking for the coastal cities and the inland cities.

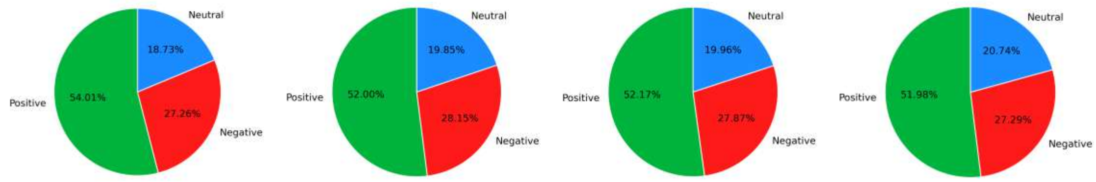
City	Country	By the Sea?	Compound Score	Ranking
Amsterdam	Netherlands	yes	0,1445	1
Berlin	Germany	no	0,1475	2
Madrid	Spain	no	0,1483	3
Lisbon	Portugal	yes	0,1562	4
London	United Kingdom	no	0,1598	5
Brussels	Belgium	no	0,1677	6
Oslo	Norway	yes	0,1683	7
Paris	France	no	0,1794	8
Helsinki	Finland	yes	0,1939	9
Dublin	Ireland	yes	0,2047	10
Athens	Greece	yes	0,2053	11
Rome	Italy	no	0,2110	12
Stockholm	Sweden	yes	0,2153	13
Copenhagen	Denmark	yes	0,2289	14
Vienna	Austria	no	0,2312	15
Luxembourg	Luxembourg	no	0,2549	16
Mean values coast			0,1896	8,625
Mean values inland			0,1875	8,375

Table 6.8: Ranking by compound score per tweets of all 16 cities from negative to positive

This table shows which cities have the most negative or the most positive compound score. In the top 3 of the most negative scores are Amsterdam, Berlin and Madrid. After these top 3, a small gap to the next city can be seen. Also in the other direction, namely the top 3 cities with the most positive values, a gap between the top 3 and the next cities can be seen. The cities with the most positive compound score are Luxembourg, Vienna and Copenhagen. In both categories, the distribution of cities is equal between coastal and inland. Two cities each are inland and one city each is on the coast.

The ranking shows a good mixture of the two groups, which can be seen in the close result for the average values in the ranking and the average compound score. On the coast, the range of scores is from 0,1445 (Amsterdam) to 0,2289 (Copenhagen) with an average score of 0,1896. The range is similar in the inland with scores from 0,1475 (Berlin) to 0,2549 (Luxembourg) with an average score of 0.1875. The difference in the respective average compound scores is 0,0021, a very small difference. In the inland, the lower values of the compound score can be observed, whereas in the ranking, the cities on the coast tend to be more negative.

If one looks at the top 3 cities with the most negative compound score as pie charts compared to the pie chart across all cities, see figure 6.12, one can see that the share of the positive category decreases for the three cities. This value is 54,01% across all cities, but for the three cities the value is only about 52%. In the other two categories, therefore, the proportion has increased. There are more neutral as well as negative (though rather close) tweets than in the average of all cities.



(a) Overall (b) Amsterdam (c) Berlin (d) Madrid

Figure 6.12: Overall classification and classification of the three cities with the most negative compound score

In figure 6.13 a comparison of the average compound scores of the two groups is shown over the two years 2020 and 2021.

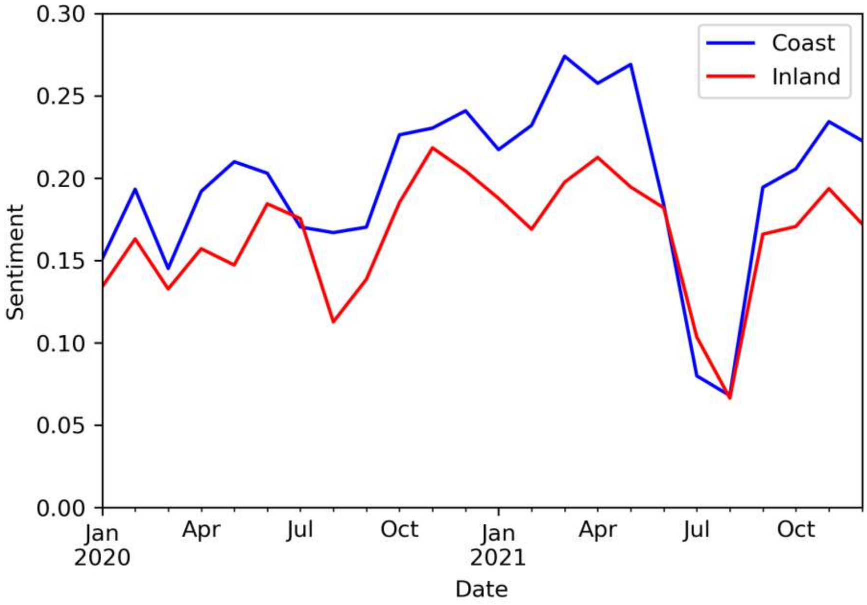


Figure 6.13: Average compound scores in the months of 2020 and 2021 per group

Again, similarities between the two groups are apparent, with cities on the coast having slightly more positive scores across the months. In both groups, the low scores can be seen in the summer months, although for cities on the coast August in 2020 does not have the most negative score over that year, as the month of March has a even more negative score in 2020. Neither in the one nor in the other group are trends in positive or negative directions discernible, only larger fluctuations in 2021.

In summary, with a overall compound score of 0,189, a slightly positive sentiment can be seen across all cities. The differences between the cities are small and have values between 0,14 and 0,26. Also if one looks at the classification of the tweets, one sees with 54,01% a majority of positive tweets. Based on the results of the compound score, the cities with the most negative mood about climate change are Amsterdam, Berlin and Madrid. For these three cities, the results of the compound score and the classification are summarized in figure 6.9.

City	By the Sea?	Compound Score	Positive	Negative	Neutral
Amsterdam	yes	0,1445	0,5200	0,2815	0,1985
Berlin	no	0,1475	0,5217	0,2787	0,1996
Madrid	no	0,1483	0,5198	0,2729	0,2074

Table 6.9: Top 3 cities with the most negative compound score

The table shows that the cities do not belong to only one group. Also in the top 3 most positive cities, the distribution is the same with two inland cities and one coastal city. Over the whole ranking a good mixture of the two groups can be seen. The difference in the averages of the compound scores of the two groups is marginal with 0,0021. In the ranking, both groups also achieve almost the same result. From this it can be concluded that there are not only similarities to one's own group, but similarities across all cities regardless of the group. These groups do not differ in the comparison of the sentiment.

CHAPTER 7

Deployment

In this phase, the results of the polarizing words and sentiment analysis are summarized visually. For this purpose, a map of Europe is shown, in which the levels of similarity and sentiment can be seen.

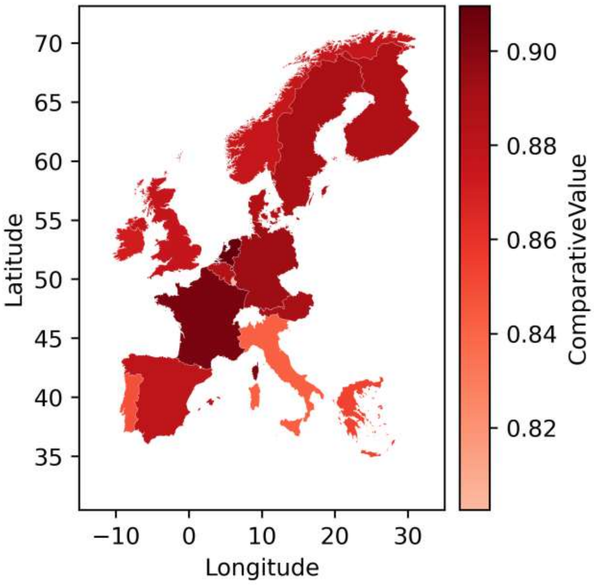


Figure 7.1: Map of Europe with the levels of similarities of polarizing words

In figure 7.1 the countries of the respective capitals can be seen, which were analyzed in the previous phases. This allows the map of Europe to visually show in which capitals there are high similarities to other capitals. Countries with high similarity, i.e. a high comparative value, are colored darker than countries with lower similarities. The underlying data are the comparative values from figure 6.7 in the Evaluation phase. There one could already see that Amsterdam and Paris have the highest similarities to other cities, therefore the Netherlands and France are the darkest colored countries. On the other hand, Italy, Portugal and especially Luxembourg are the lightest, since Rome, Lisbon and Luxembourg had the lowest similarities.

In figure 7.2 the levels by sentiment can be seen. The underlying data for this can be found in table 6.8 in the Evaluation phase. The top 3 cities with the most negative sentiment can be seen well, as these are the countries shown darkest. On the other hand, the countries Luxembourg, Austria and Denmark are colored the lightest and represent the capitals with the most positive sentiment.

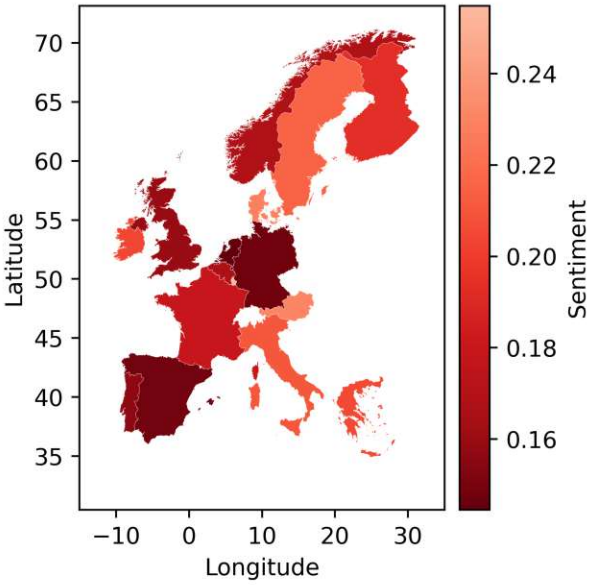


Figure 7.2: Map of Europe with the levels of sentiments

CHAPTER 8

Conclusion

The target of this work was to gather tweets in the context of climate change and information about cities to join them based on the user profile location of Twitter users to then identify the current mood and polarizing words in the context of climate change on a city level. A comparison within coastal cities and within inland cities aimed to show whether there were similarities in these groups and whether these groups differ from each other. For this purpose, data mining was performed using the CRISP-DM methodology. In the individual phases, the data was reviewed, analyzed, processed and evaluated. In order to answer the research questions, tweets in the period from 01.01.2020 to 31.12.2021 were collected in English language, which were subtracted based on certain hashtags. The selection of hashtags was analyzed and done with the help of already made analyses of hashtags in the field of climate change from the literature. The choice was made for more general hashtags, rather than specific topics.

In the analysis for the assignment of tweets to European capitals, it was found that significantly more tweets (8,62% of all tweets) can be assigned to European capitals using the user profile location than using GPS data (0,09% of all tweets). Therefore, a separate city dataset with the European capitals was created, which contained the European capitals in their official languages and in English, and provided information about whether the cities are located on the coast or inland. After an analysis of the values and challenges within the user profile location, an assignment of the tweets to the city dataset was done with the help of defined characteristics from analysis. For these assigned tweets, disturbing factors were removed and a similarity cleaning was done to filter out possible bots or similar tweets from users. This resulted in a dataset of 259.737 tweets that were matched to European capitals. The top 16 cities with the most tweets were selected because they had a high number of tweets and an equal distribution of coastal and inland cities (eight to eight).

Using VADER and an adapted TF-IDF, the data on sentiment and polarizing words were modeled and analyzed. The results show that for the polarizing words more similarities were found within their own groups. Similarities were compared using three methods: top 20 words, words above the 97,5th percentile and nouns above the 97,5th percentile. The total score across the three methods displays that 78 comparisons show more similarities to the own group and 18 comparisons show more similarities to the other group. Thus, the European capitals show significant more similarities to cities in the same group than to cities in the other group. Across all cities, the words with the most occurrences are 'climate', 'change', 'need', 'new' and 'world'. In contrast, when evaluating the sentiment scores, there were not only similarities to one's own group, but similarities across all cities regardless of the group. Thus, the groups do not differ in the comparison of the sentiment. All cities have a positive compound score, with an average value of 0,189. The three cities with the most negative compound score are Amsterdam, Berlin and Madrid. From these results, it can be concluded that there is a difference in the tweeting behavior of users in the different groups of cities. While the general sentiment was similar across the groups, there was a significant difference in tweets when it came to polarizing words.

With this work it was shown that there are other possibilities to assign tweets to cities than GPS data. This way, it was possible to assign significantly more tweets and thus extract significantly more information from the tweets. The selected method of assigning the data based on the user profile location shows new options that can be further optimized in the future. Especially in regions where there are significantly fewer tweets and analyses are to be made, this can bring added value. The comparisons on a city level show that there are different topics in the different cities that occupy the population there. This can create a picture of what words polarize and what the mood is over a certain period of time. For this thesis, the topic of climate change was addressed, but using a different set of hashtags, the same analysis can be done in other areas as well.

8.1 Limitations

In this work, only data from Twitter was used and no data from other social networks. Other social networks are mostly more image or video driven and therefore do not offer as many possibilities when analyzing texts in a certain area. Twitter is a social network that is mainly text driven. However, when analyzing Twitter data, there may not be a representative opinion of humanity across all social classes and age structures. There may be a certain lack of distribution. But in general Twitter is a platform where one can get a good overview of different opinions in trending topics.

Since everything on Twitter is text-driven and information about users can also be filled in using free text, it is difficult to fully automate the assignment of users using user profile location. This is because many factors would have to be taken into account and all possible scenarios that users are able to enter would have to be covered. Therefore, it is necessary to perform manual checks and research to make the assignment as accurate as possible. It would be possible to include other factors like GPS or tweet content in the assignment or in machine learning processes to get more information about the user or a more accurate assignment. However, this was not part of this work.

Free texts enable users to specify several cities in the user profile location and not just exactly one city. As a result, some users specify different cities in enumerations, which directly or indirectly have an influence on the analysis in this work. For example, several European capitals may be specified, resulting in tweets that fell into the evaluations of several cities as well. Or cities outside of Europe, which led to some tweets referring to events outside of Europe.

Another limitation is that only tweets in English were analyzed in this thesis and thus many opinions and tweets in various official languages are not taken into account. When selecting the cities with the most tweets, a clear west-east divide can be seen, which can also be seen in the maps of Europe in the Deployment phase. The cities selected for evaluation are located in Western, Central and Northern Europe. The west-east divide may not necessarily be due to the restriction to the English language, there may also be other reasons such as the use of Twitter or the level of interest in climate change in eastern countries of Europe.

List of Figures

List of Tables

Bibliography

[ALG+16] Beatrice Alex, Clare Llewellyn, Claire Grover, Jon Oberlander, and Richard Tobin. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, pages 3936–3944. European Language Resources Association (ELRA), 2016.

[APA+17] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *CoRR*, abs/1707.02919, 2017.

[BCN20] Omar Benjelloun, Shiyu Chen, and Natasha F. Noy. Google dataset search by the numbers. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, volume 12507 of *Lecture Notes in Computer Science*, pages 667–682. Springer, 2020.

[BKJ19] Venkateswarlu Bonta, Nandhini Kumaresh, and N. Janardhan. A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian Journal of Computer Science and Technology*, 8(S2):1–6, 2019.

[Bro20] Katherine Brown. 2020 Tied for Warmest Year on Record, NASA Analysis Shows, 2020. Available at: `https://www.nasa.gov/press-release/2020-tied-for-warmest-year-on-record-nasa-analysis-shows` (Last accessed on 2022-05-23).

[CCK+00] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. CRISP-DM 1.0: Step-by-step data mining guide. SPSS, 2000.

[CP21] Jonnathan Carvalho and Alexandre Plastino. On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artificial Intelligence Review*, 54(3):1887–1936, 2021.

[CRM+15] Emily M. Cody, Andrew J. Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth. Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLoS ONE*, 10(8):1–18, 2015.

[DKL19] Biraj Dahal, Sathish A.P. Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1):1–20, 2019.

[EKSA22] Dimitrios Effrosynidis, Alexandros I. Karasakalidis, Georgios Sylaios, and Avi Arampatzis. The climate change Twitter dataset. *Expert Systems with Applications*, 204(May):117541, 2022.

[HG14] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*, pages 216–225. The AAAI Press, 2014.

[HLLE14] Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. Word cloud explorer: Text analytics based on word clouds. In *47th Hawaii International Conference on System Sciences, HICSS 2014, Waikoloa, HI, USA, January 6-9, 2014*, pages 1833–1842. IEEE Computer Society, 2014.

[HMB17] B S Harish and Revanasiddappa M B. A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents. *International Journal of Computer Applications*, 164(8):1–7, 2017.

[HZ15] Ahmed Abdeen Hamed and Asim Zia. Mining climate change awareness on twitter: A pagerank network analysis method. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Marina L. Gavrilova, Ana Maria Alves Coutinho Rocha, Carmelo Maria Torre, David Taniar, and Bernady O. Apduhan, editors, *Computational Science and Its Applications - ICCSA 2015 - 15th International Conference, Banff, AB, Canada, June 22-25, 2015, Proceedings, Part I*, volume 9155 of *Lecture Notes in Computer Science*, pages 16–31. Springer, 2015.

[KKP+21] Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi. Analysis of geotagging behavior: Do geotagged users represent the twitter population? *ISPRS International Journal of Geo-Information*, 10(6), 2021.

[KM06] Lukasz A. Kurgan and Petr Musilek. A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review*, 21(1):1–24, 2006.

[KS14] Andrei P. Kirilenko and Svetlana O. Stepchenkova. Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, 26(1):171–182, 2014.

[Kum21] Prashant Kumar. Climate Change and Cities: Challenges Ahead. *Frontiers in Sustainable Cities*, 3(February):1–8, 2021.

[L$^+$66] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.

[LM22] Mitchell Lennan and Elisa Morgera. The Glasgow Climate Conference (COP26). *International Journal of Marine and Coastal Law*, 37(1):137–151, 2022.

[LRK16] Farhad Laylavi, Abbas Rajabifard, and Mohsen Kalantari. A Multi-Element Approach to Location Inference of Twitter: A Case for Emergency Response. *ISPRS International Journal of Geo-Information*, 5(5):56, apr 2016.

[MIB21] Sajib Mandal, Md Sirajul Islam, and Md Haider Ali Biswas. Modeling the potential impact of climate change on living beings near coastal areas. *Modeling Earth Systems and Environment*, 7(3):1783–1796, 2021.

[MPCOF$^+$21] Fernando Martinez-Plumed, Lidia Contreras-Ochando, Cesar Ferri, Jose Hernandez-Orallo, Meelis Kull, Nicolas Lachiche, Maria Jose Ramirez-Quintana, and Peter Flach. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061, 2021.

[MRHC21] Dwiny Meidelfi, Indri Rahmayuni, Taufik Hidayat, and Dikky Chandra. TF-IDF Implementation for Similarity Checker on The Final Project Title. *International Journal of Advanced Science Computing and Engineering*, 3(1):40–52, 2021.

[MSM22] Nabila Mohamad Sham and Azlinah Mohamed. Climate change sentiment analysis using lexicon, machine learning and hybrid approaches. *Sustainability*, 14(8):4723, 2022.

[MSMFB09] Oscar Marbán, Javier Segovia, Ernestina Menasalvas, and Covadonga Fernández-Baizán. Toward data mining engineering: A software engineering approach. *Information Systems*, 34(1):87–107, 2009.

[QA18] Shahzad Qaiser and Ramsha Ali. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 181(1):25–29, 2018.

[RAP+13] Erica C. Rodrigues, Renato M. Assunção, Gisele L. Pappa, Renato Miranda, and Wagner Meira Jr. Uncovering the location of twitter users. In *Brazilian Conference on Intelligent Systems, BRACIS 2013, Fortaleza, CE, Brazil, October 19-24, 2013*, pages 237–241. IEEE Computer Society, 2013.

[Sal21] Jeffrey S. Saltz. CRISP-DM for data science: Strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, December 15-18, 2021*, pages 2337–2344. IEEE, 2021.

[SFW+20] Wen Shi, Haohuan Fu, Peinan Wang, Changfeng Chen, and Jie Xiong. #Climatechange vs. #Globalwarming: Characterizing two competing climate discourses on twitter with semantic network and temporal analyses. *International Journal of Environmental Research and Public Health*, 17(3), 2020.

[SKG21] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181(2019):526–534, 2021.

[SKMM20] Siddhartha B S, Divya Khyani, Niveditha N M, and Divya B M. An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Journal of University of Shanghai for Science and Technology*, 22(10):350–357, 2020.

[SM19] Kotagiri Srividya and A. Mary Sowjanya. Aspect based sentiment analysis using POS tagging and TFIDF. *International Journal of Engineering and Advanced Technology*, 8(6):1960–1963, 2019.

[STC21] Stephan Schlosser, Daniele Toninelli, and Michela Cameletti. Comparing methods to collect and geolocate tweets in Great Britain. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(1):1–20, 2021.

[TBCS20] Aman Tyagi, Matthew Babcock, Kathleen M. Carley, and Douglas C. Sicker. Polarizing tweets on climate change. In *Social, Cultural, and Behavioral Modeling - 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18-21, 2020, Proceedings*, volume 12268 of *Lecture Notes in Computer Science*, pages 107–117. Springer, 2020.

[WH00] Rüdiger Wirth and Jochen Hipp. CRISP-DM: towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, Manchester, UK, April 11–13, 2000*, pages 29–39, 2000.

[ZHS18] Xin Zheng, Jialong Han, and Aixin Sun. A Survey of Location Prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018.