



**Publication series of Professorship for Global Supply chain
Management**

Chair Holder: Prof. Dr. Aseem Kinra

**Big Data Analytics for controlling supply chain
performance**

Riemer, Torben; Warnke, Phillip

Year 2020

Supervisor: Prof. Dr. Assem Kinra | Debarshee Bhardwaj

Table of Contents

Table of Contents	II
List of Figures	IV
List of Tables.....	V
List of Abbreviations.....	Error! Bookmark not defined.
1 Introduction	1
1.1 Problem statement	1
1.2 Aim of the project.....	3
1.3 Stakeholders.....	3
2 Methodology	5
2.1 Literature research and findings	5
2.2 Literature review.....	5
2.2.1 Environments and goals in transport logistics.....	6
2.2.2 Use of google maps in transportation.....	7
2.2.3 Use of google maps in general traffic	8
2.2.4 Predictive analytics	9
2.2.5 How has predictive analytics been used before?.....	10
2.2.6 Text analytics	11
2.2.7 How has text analytics been used before?.....	13
2.3 Using traffic reports in transport logistics	14
3 Analysis.....	18
3.1 Document selection and structure.....	18
3.2 Classification of traffic reports	19
3.3 Counting of traffic reports	22

3.4	Output of framework	23
4	Conclusion.....	25
	References	27
	Appendix	30

List of Figures

Figure 1: Total length of traffic jams	1
Figure 2: Volume of road transportation	2
Figure 3: Stakeholder Groups	4
Figure 4: Linear Regression between traffic volume and Popular times	9
Figure 5: Linguistic Foundation of Text analytics	11
Figure 6: Natural Language Processing	12
Figure 7: Text mining framework for traffic reports.....	13
Figure 8: Example of a traffic report for text processing	16
Figure 9: Filter options for traffic reports	18
Figure 10: Structure of traffic reports	19
Figure 11: Classifying of traffic reports	20
Figure 12: Example of a traffic report for traffic jam	21
Figure 13: Example of a traffic report for slow-moving traffic	21
Figure 14: Example of a traffic report for danger	22
Figure 15: Example of a traffic report for neglected reports	22
Figure 16: Traffic report 01.01.2016 Route Hamburg – Köln	24

List of Tables

Table 1: Manual extracted word list.....	16
--	----

1 Introduction

Using trucks in the supply chain to transport goods from e.g. the supplier to the manufacturing step produces a lot of challenges. Big risks for the timetable in the supply chain are trucks, stuck in traffic jams.

The “ADAC Staubilanz 2018” measures every traffic jam for every year in Germany. This means in numbers 745.000 traffic jams and 1.528.000 kilometers total length of all traffic jams in 2018. In comparison with 2017, this means growth of 3% for the number of traffic jams and 5% growth for the length of traffic jams (“Autobahnen in Deutschland - Gesamte Staulänge bis 2018,” 09.12.2019)

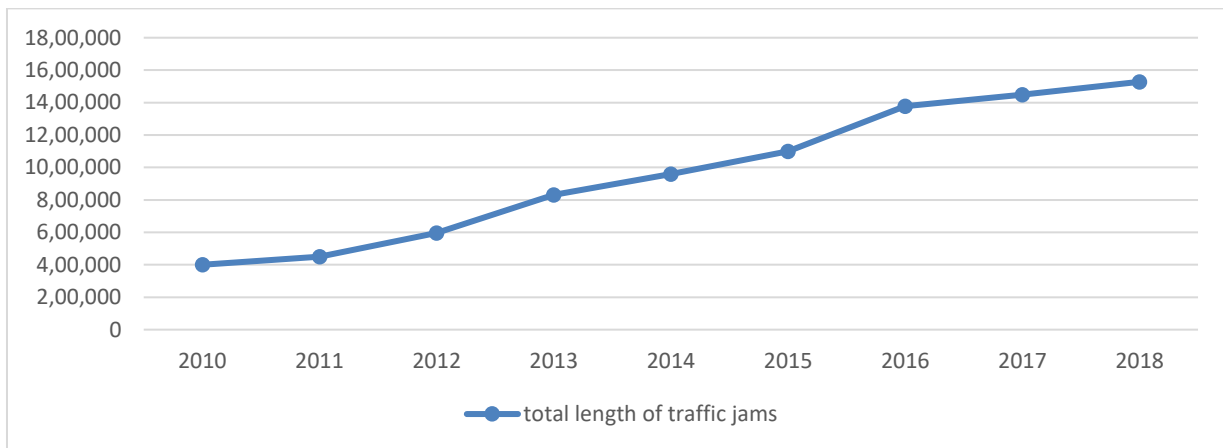


Figure 1: Total length of traffic jams (ADAC 2018)

1.1 Problem statement

The use of trucks for the transport of goods produces a lot of challenges for the company, resulting in risks for the timetable in the supply chain.

Several statistics show the increasing flow of goods in the logistics network, the growing traffic, the length of traffic jams and the time that is being lost during these jams (“Staustunden auf deutschen Autobahnen,” 2020)

The volume of goods is increasing in Germany. Especially on the road, the amount of goods has been increased from 3402,5 in 2011 to 3746,6 in 2018 (in million tons), which is an increment of 10,1%. In comparison to inland shipment, where in the same amount of time was a decrease meant of 10,9% and the transport on the rail, which also has decreased by 4,9%. The increase on the road seems even bigger or a redistribution has been taken place between these forms of transportation. Actual forecasts are attempting an additional increase in street

transportation of 7% until 2022 (“Güteraufkommen in Deutschland je Verkehrsträger,” 09.12.2019)

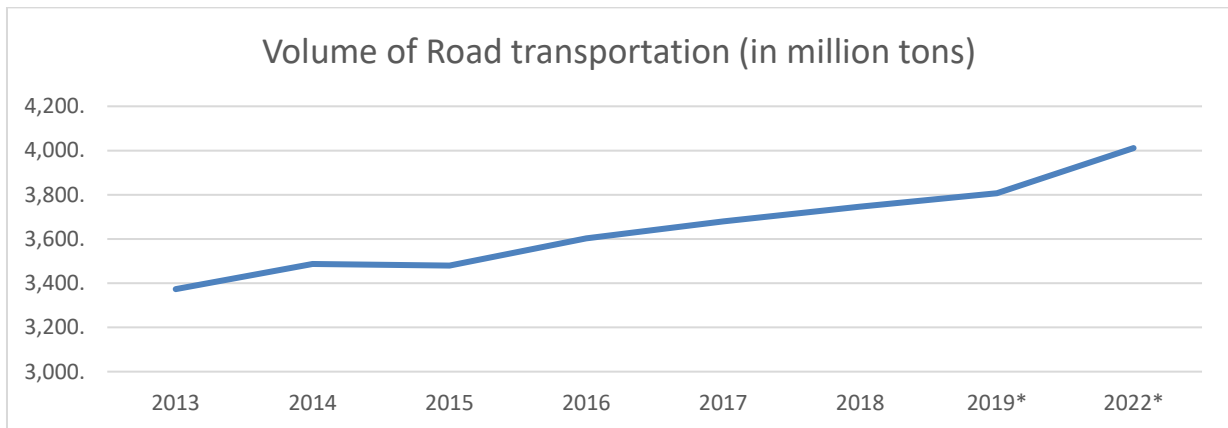


Figure 2: Volume of road transportation (Transportaufkommen im deutschen Straßenverkehr bis 2022, 2019)

Actual statistics show an increase from 450.000 kilometers in 2011 to 1.528.000 kilometers in 2018, which is an increase of 340% in 7 years. In 2018 that meant a total delay of 459.000 hours related to traffic jams (Autobahnen in Deutschland - Gesamte Staulänge bis 2018, 2019)

Coming from the topic: “big data analytics for controlling supply chain performance”, this working paper is about improving transport routes by using automated predictive analysis of traffic reports.

Congestion delay produces logistical consequences for the supply chain performance for stored goods. A delayed delivery produces disruption at the next level in the supply chain. Possible consequences can be missed booking-in time, which means to wait until the next available slot or the delivery must be rejected (McKinnon, 1999, p. 19).

As a result, the logistics companies are dependant on using the road network for transport purposes, this traffic problem becomes a problem for the entire transport logistics industry. To deal with this problem, there have been approaches in the past that rely on the use of Google data, in particular Google Maps data. These approaches are explained in this project and treated further as a common problem case is the lack of available data sources.

In summary, the rising volume of traffic in combination with an increasing volume of freight traffic is leading to ever-longer traffic jams on German roads. This, in combination with the inadequate use of text-based data sources, which can contribute to the expanse of existing solution approaches based on google maps, leads to the following research question.

The following research question is going to be answered throughout the working paper:

Can google maps be used for transportation problems during the supply chain?

How can historic traffic reports be used to predict future traffic?

1.2 Aim of the project

The overall aim of the project is to deny trucks losing time in traffic jams. Meaning optimizing their driven time windows for delivery, regarding stuck traffic on the street.

The improvement should result in a steady flow of goods and more accurate planning with deliveries.

To be able to show the best time windows, we will be using text analytics and machine learning to have an opportunity to examine the big amount of data, that comes from traffic reports. This will lead the project to an overall view of traffic jams according to historic and actual data. This will make it easier and more comfortable for companies to plan their transport routes properly.

It's important to use text-based data sources, even if there are already databases for traffic jams because databases are just about 5% of the total amount of available data sources. Other types of sources are images, audio, and video, which are more often available for companies (Kinra et al., 2019, p. 174).

The overall amount of text-based data is growing even more in the future, according to social media usage and the globalization (Kroker, 2018). Whether there are Facebook groups to tell other groupmates about traffic jams, several apps to warn early for accidents or Twitter accounts from organizations, that work with traffic data or to avoid traffic problems like the ADAC for example.

The project work identifies gaps in actual prediction of vehicle routing. To handle the identified gaps and become more precise in traffic prediction this study presents the potential in gaining information from historic traffic reports.

1.3 Stakeholders

To explain possible benefits from this working paper, it is needed to determine the Stakeholders, who are inflicted by transport logistics. The main relation that is needed to be expected is the relation between suppliers and receivers because this relationship has a major impact on how a logistics company works, as it is necessary to fulfill the receiver's demand (Estrada and Roca-Riu, 2017, p. 167).

In this environment, suppliers and receivers can be defined in multiple ways, as the receivers can be other companies in a business to business relation, or an end consumer in a business to consumer relationship. Either way, the supplier can be another producing company, which is delivering its goods as a supplier to a further processing company or a logistics service provider.

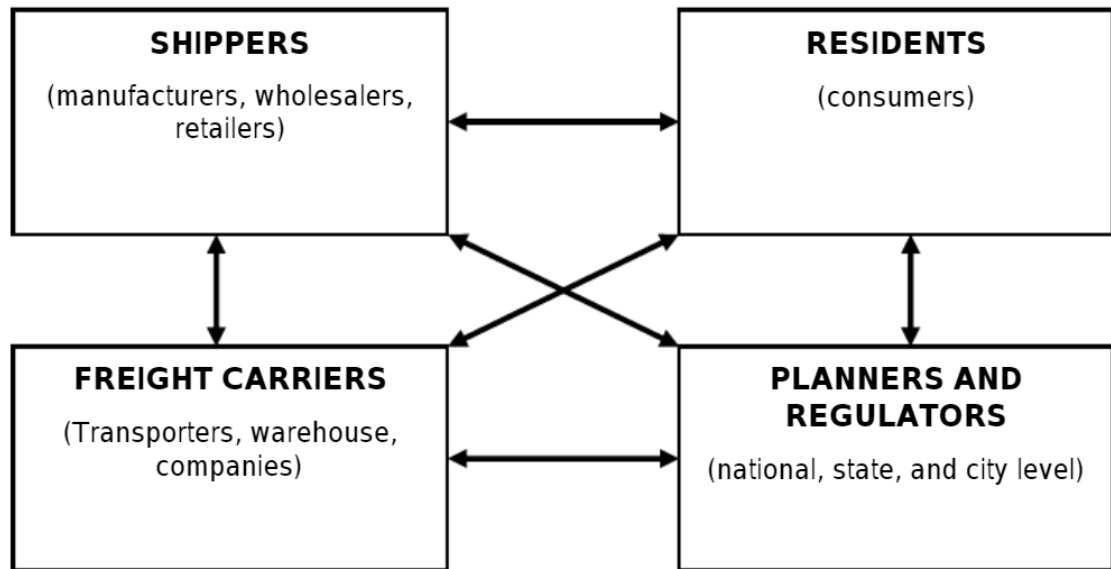


Figure 3: Stakeholder Groups (Taylor, 2005, p. 5)

Figure 3 shows the groups of stakeholders of urban transport logistics. The stakeholders can be separated into Shippers, which symbolize the upstream part of the supply chain and the Residents as the downstream end of the supply chain. Freight Carriers are stakeholders in any transportation issue, while the Planners and Regulators mainly belong to the environments of logistics, which will be discussed later, but also take their part as stakeholder by benefitting from positive externalities caused by transport optimization such as reduced pollution, noise and the development of transportation infrastructure. Each of these stakeholders has its own objectives to the topic, while Planers and Regulators set the framework for logistics to take place in, the Freight Carriers have to accept certain rules to fulfill their customers' needs, like driving times for example (Taylor, 2005, p. 4f.).

The most important stakeholders for this project are the shippers and the freight carriers, as these are the stakeholder groups that are most likely to benefit from an improvement in the prediction of traffic jams and are able to provide large amounts of text-based data on traffic reports.

2 Methodology

2.1 Literature research and findings

For finding relevant literature for this work, it was necessary to search through multiple search engines, including mainly the Website from the Bremen State and University Library, which led to further databases of literature like Elsevier, Researchgate, Wiley online library, Statista, and Springer. For further research, the references of important literature for this paper have been used, to get more detailed information about the topics. As well, the Google search engine has been used, to find relevant literature and companies that are using predictive analytics, text analytics or google based functions in their everyday work to give some examples. To find specific literature, it was needed to search for specific keywords, as logistics, transportation and big data analytics are taking part in multiple areas, which are not all-important for this paper. In the following is a presentation of the list of keywords that have been used to generate this paper:

List of keywords:

- Logistik (logistics)
- Transport (transport)
- Google Maps
- Popular Times
- Transportation delay
- Stau (traffic jams)
- Predictive analytics
- Text analytics
- Predictive analytics history
- Text analytics history
- Text mining techniques
- Data mining
- Text mining
- Volume of transportation
- Length of traffic jams
- Stakeholders of transport logistics
- Growth of transport volume

As well as special combinations of the keywords mentioned before, to get more specialized literature, especially:

- Using google maps in transport logistics
- Using Google Popular times in transport logistics
- Using text analytics in transport logistics
- Using predictive analytics in transport logistics

2.2 Literature review

In the following Chapters, an overview of the topics covered in this paper will be given. This includes the environments of logistics, the use of google maps in general traffic problems, as well as in general transportation issues. Furthermore, the principles of predictive analytics, as well as some examples of how predictive analytics has already been used will be explained. Besides that, an introduction to text analytics will be given, as well as examples for the usage of text analytics in the past. Finally, methods of text modeling will be shown which provide the foundation for the framework of this working paper.

2.2.1 Environments and goals in transport logistics

Transport logistics is being influenced by logistics environments. Those logistical environments can be spread into 4 different types including the community, nature, technology, and economy. Each of these environments has its traits, which influence the considered problem of traffic jams in transport logistics, which will be discussed. According to Ulrich, the community is regulating governmental rights and laws like the total weight of the transport, the time, that one driver can drive without pause and driving bans. The Community aspect also includes public infrastructure like freeways, bridges and parking lots (Barwig, 2013, p. 58 ff.).

The aspect of nature includes the used resources, the climate, its changing and the environment itself. In the case of transport logistics, it is mostly about the usage of fuel and its impact on climate change, which is discussed in today's plans for optimization. Using predictive analytics to reduce the duration of transports can also help saving fuel, which saves money for the company and CO² for the climate (Barwig, 2013, p. 59 f.).

The third aspect, “technologies” is where predictive analytics takes place. It also includes traffic guidance systems, communication- and information systems, as well as planning- and structurization actions. Especially the information technologies are growing and can take a major impact on optimizing logistics solutions (Barwig, 2013, p. 60 f.).

With the economy as the last aspect, there is the foundation for any optimization process, according to the structures of communication, like GPS- and tracking systems, which are used to plan the vehicle routing (Barwig, 2013, p. 60 f.).

These systems are also able to produce a big amount of actual, structured traffic data, which will be discussed later.

The discussed need to be taken into concern while defined the goal of transport logistics.

Transport logistics is a partial discipline of the whole logistics network to show, which goals are important in transportation, it is first necessary to keep in mind, that transport logistics is used in every logistics process. The logistics processes can be divided into procurement, production, distribution, and disposal. The major goals of transport logistics can be divided into cost goals and performance goals. Cost goals include the reduction of costs in transportation,

warehousing, transshipment, and packaging. Also, the performance goals are in general to improve the delivery time, reliability, flexibility, and quality. It is also a performance logistics goal to improve the informativeness, which can profit by using predictive analytics (Barwig, 2013, p. 68 f.).

General conditions, growing costs as well as growing performance goals are major reasons for the increase of the complexity in transport logistics networks (Barwig, 2013, p. 70).

As an example, this can be shown in decreasing average shipment size, divergent delivery time, increasing the need for flexibility to answer, order modifications and an increasing number of traffic disruptions.

2.2.2 Use of google maps in transportation

The usage of simple google maps can not be recommended for using supply chain and transportation issues, because these issues are more than just the simple shortest route between point A and point B. But the huge amount of Data that google is collecting every day from millions of mobile phones, including positions, movement speed and the density of data points in specific areas makes google maps the go-to application for a personal route (Mangal, 2020)

Due to a lack of historical data, it seems not to be predictable for google maps if there could be traffic disruption on the route or other issues that don't benefit a precise prediction of travel time. Especially on long distances, which includes most shipments, conditions like traffic issues are getting more important to be predicted early to choose another route. This means, that a workable program for optimizing transport logistics problems needs to be more precise and have the ability to display some form of dynamic route optimization. And those are only issues considering one truck's routing problem. When it comes to a bigger fleet of logistic service providers, the capacity that is given by google maps is simply not enough (Mangal, 2020)

Thereby, other companies recognized this problem and the potential of google maps and it's a huge database to build up their own systems and applications by implementing maps as one part of their solution. In the following there will be some examples of apps and companies that are provided, using implemented google maps to face supply chain and transportation optimization.

i. **Ubilabs**

Ubilabs is a german company located in Hamburg, who provides their customer's solutions using google data, including google maps like location monitoring and navigation ("Google Maps – Applications & IT Integration for companies | Ubilabs,"15.12.2020).

ii. **Route Savvy**

Route Savvy is a program and application that uses google maps data to provide their customers a solution for planning routes, waypoints and provides a GPS tracker for smartphones to get an overview of the fleet's actual location. Special benefits are the options

to show multiple routes at the same time and the optimization of hundreds of routes ("Powerful, Affordable Route Planner | RouteSavvy.com," 15.12.2020)

iii. **Tailwind**

Tailwind is a web-based trucking and freight broker software, which includes many important tasks to transportation, as freight brokerage, mileage calculations and others with the opportunity of connecting key partner integrations ("Enterprise Transportation Management Software - Products | Tailwind TMS," 15.12.2020).

Concluding the usage of google maps in transportation, it is getting visible, that the functions provided by google not seem to be individual enough for solving complex transportation or supply chain issues, but the usage of the information that is distributed by google can benefit in multiple ways. The given examples are showing that it's necessary to link more information with google maps to produce usage for supply chain management.

2.2.3 Use of google maps in general traffic

Can Google Maps Popular Times be an alternative source of information to estimate traffic-related impacts? Is the title of research by Pavlos Tafidis and others, which aim it is to show, which impact the analysis of popular times on google can have in reducing traffic volumes, Co² Emissions and other externalities produced by transport logistics (Tafidis et al., 2018).

„In the last years, many studies have shown attention to explore the potential of using web-based data sources for transport planning, management or operation. The real-time information that they provide allows commuters to improve their travel experience and transportation authorities to enhance the quality of their services. More specifically, it can allow city and transport planners to gain a better understanding of mobility patterns and needs, while for individuals to move freely and reducing travel time.“ (Tafidis et al., 2018, p. 3)

This quote shows, what the source is about to explain and what kind of information it will provide to be worked on in this project. The methodological approach of that study is, to see relationships between the google „popular times“ as a prediction tool with traffic volumes, emissions and travel time.

Furthermore, the study divides the passing vehicles by categories for passenger cars, medium-heavy vehicles, and heavy-duty vehicles by using videotaping. It also uses GPS data to see the average speed (Tafidis et al., 2018, p. 6).

The study is regarding the correlation between popular times and the traffic volume and it identifies strong results. The comparison is made between the Popular times' table of google maps (figure) and the traffic volume per 15 minutes. The Popular times' table is set for a value between 0 and 1 (Tafidis et al., 2018, p. 7f.).

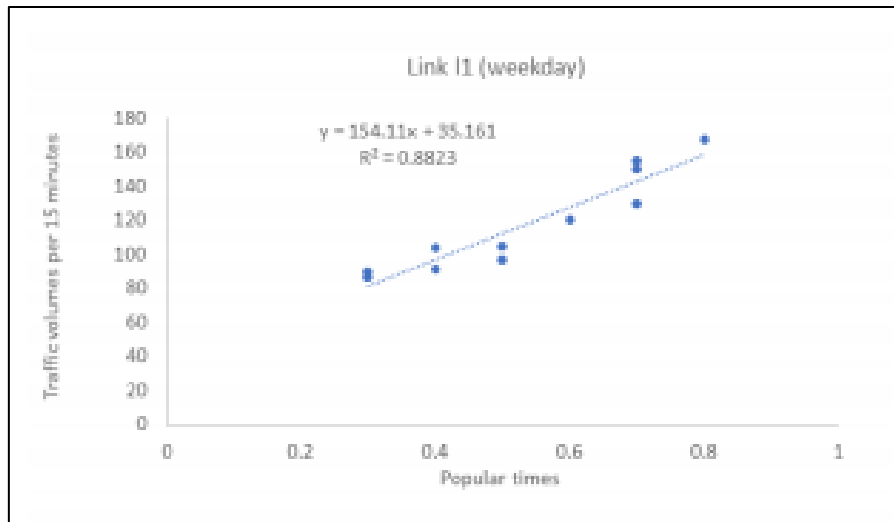


Figure 4: Linear Regression between traffic volume and Popular times (Tafidis et al., 2018, p. 9)

The study concludes, that the Google Maps Popular times is usable to predict traffic-related impacts under a set of restrictions, but it is necessary to collect more data or use data sets to be able to predict more precisely (Tafidis et al., 2018, p. 14)

Before the implementation of a monitoring system based on this type of information, it will be necessary to collect data over an extended period or to have access to similar data sets, e.g. based on absolute values for urban zones on open data platforms. (Tafidis et al., 2018, p. 14)

2.2.4 Predictive analytics

What is predictive analytics? – An important question to be answered, before explaining text analytics as a partial discipline of it. In the core, predictive analytics is a technology, that learns from data to predict things in the future. In general, these “things” can be for example the behavior of consumers, the flight risk of new employees, the decision making process in politics, the weather or as in this working paper the prediction of traffic jams and other issues on roads to optimize transportation logistics (Siegel, 2016, p. 8f.).

Predictive analytics can be used to analyze structured and unstructured forms of data like charts, tables, texts, videos, and anything, that can be used to feed a machine to get information out of it. The more unstructured a source of data is, the more pre-processing is needed to be done to get it in the right shape for a program to work on it. Explained simply: if you are reading an article in your hometown newspaper, you must read hundreds of words to finish it. But only a few of them are important to the text's core value, which can be dates, places, persons and interactions. Depending on the wanted output, the predictive analytics techniques, which will be discussed later are needed to set the right to filter the right information.

By collecting huge amounts of data, structuring them and letting them pass through predictive analytics systems, it is going to be possible to drive person decisions empirically (Siegel, 2016). By understanding the different influences from input to output and from output on input before even starting to decide can reduce failure and optimize planning.

The huge amount of data collected worldwide in companies is a chance of using machine learning to get better results in every asset it takes place in. Companies must learn to grow and get better results, such as a normal person must learn, too. Regarding this fact a company, that doesn't use any sort of predictive analytics can be compared to a person with a photographic memory, who decides not to think (Siegel, 2016, p. 17)

2.2.5 How has predictive analytics been used before?

Predictive analytics enables a company to get to know their customers better, predict future needs and desires based on identified trends and patterns. Predictive analytics uses the wealth of available data of a given situation with the collected data from the past, trying to identify recurring and logical trends. This way, innovative and sustainable technologies, products and services can be used to create customer's value. (Heggenberger and Mayer, 2018, p. 2f.)

Getting better knowledge about customers made the company's work stronger on prediction aspects, to be able to approach customers more targeted, which leads to a more satisfied and loyal customer. Therefore it becomes elementary for companies to make strategic and profitable decisions to benefit from a better market positioning (Heggenberger and Mayer, 2018, p. 3)

In 2003, a study by the market research company International Data Corporation came to the conclusion that the use of predictive analytics can lead to a return on investment of approximately 250%. Furthermore, the American consulting firm Nucleus calculated, that each dollar invested into predictive analytics returned by approximately 10,66\$ (Heggenberger and Mayer, 2018, p. 12)

AS an example of predicting consumer needs, the worldwide acting company Amazon is outstanding for the benefits and success of predictive analytics. Using predictive analytics to give buy recommendations led to 35% of overall sales could be traced back to this form of analytics. Another example of predictive analytics usage is the german company Otto, which uses their data to optimize their storage and procurement costs. The company's own forecast rate improved by using predictive analytics up to 40%, which means a double-digit million amount of savings in storage and procurement (Heggenberger and Mayer, 2018, p. 13)

In summary, it is clear that the use and potential of predictive analytics lead to good to very good results in many areas. It is also clear that predictive analytics can be applied wherever there is access to large databases. The use of this data can be an advantage for the company wherever collected data can be linked to specific problems or goals. The sector in which the company is active is not decisive for this.

2.2.6 Text analytics

Web and network data science: modeling techniques in predictive analytics is a book by Thomas W. Miller to show and understand the technique and usage of text analytics and other methods of predictive analytics.

Textual data in transportation research: techniques and opportunities from Aseem Kinra to show important techniques in text mining and to identify relevant sources for text analytics.

As an important part of predictive analytics, text analytics uses two primary ways to analyze texts. The bag of words approach and natural language processing are mentioned in this book. The text corpus is being analyzed to create keys, indices, expressions, and matrices, that are easier to be analyzed by a computer (Miller, 2014).

For a better understanding, Figure 5 shows the partial disciplines of how communication works. communication can be explained as a combination of grammatical rules for sentences, the words itself and the meaning expressed by it. While there are even more parts, that need to be put together for a workable source of communication, this paper is going to focus on the semantics part of communication as the meaning of words is most important to text analytics. (Miller, 2015, p. 253)

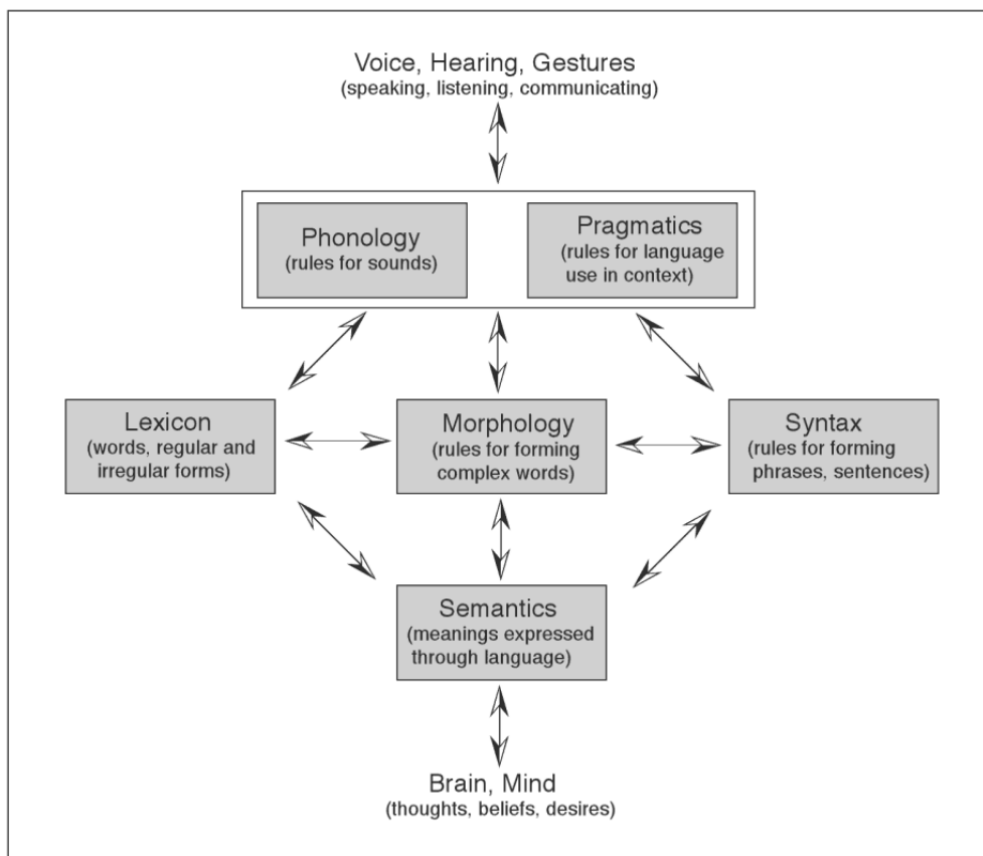


Figure 5: Linguistic Foundation of Text analytics (Miller, 2015, p. 253)

The natural language processing deals with the natural language, which is nothing more than normal speaking in the first step. In every document, that we can use for text analytics, there are sentences, which contain words. Each sentence is written in the common grammatical rules, but one meaning can be written in different types of grammar. Therefore, it is important to implement words, that are used in combination and the rules of grammar into text analytics as a foundation (Miller, 2015, p. 253ff.)

„Computer programs for natural language processing use linguistic rules to mimic human communication and convert natural language into structured text for further analysis.“ (Miller, 2015, p. 254)

It is also mentioned that earlier words in sentences are often more important than later words. As well as the words, that are written in the title of a document. On the other hand, defined and undefined articles, prepositions and pronouns do not explain the meaning of the text but are always used in texts according to grammatical rules and better understanding. These words won't take part in text analytics (Miller, 2015, p. 254).

The next step is to build word stems. As an example in the book is the word stem „market" mentioned, that also includes „marketer“, „marketeer“ and „marketing“. It is also important to have a look at words with special meaning when combined with other words. „New" and „York" were mentioned here (Miller, 2015, p. 254f.).

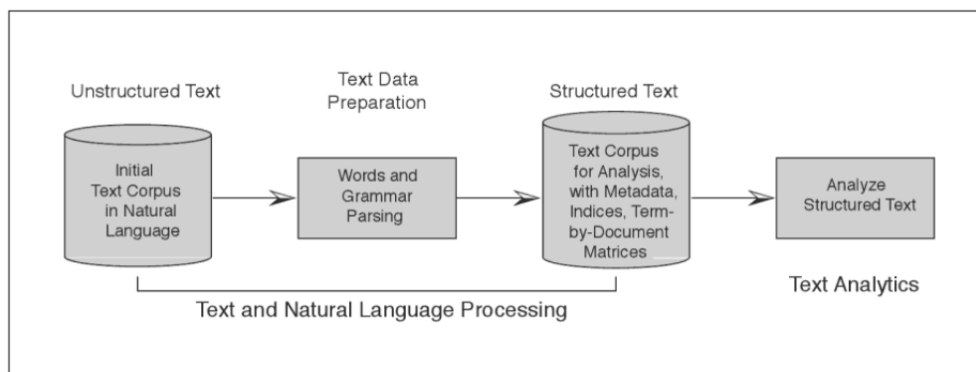


Figure 6: Natural Language Processing (Miller, 2014)

The terms-by-documents-matrix is mentioned as a key step in text analytics. The frequency of words in this matrix can be shown as an indicator of the importance of that word or word stem for the understanding of the document (Miller, 2015, p. 257).

To show what has already been done yet, there is an example from IBM. IBM's Watson is shown as what intelligent analytic programs are becoming. This program already understands the typical grammatical roles of subject, verb, object, and modifier. Also, they know nouns, verbs, adjectives and adverbs as a party of speech (Miller, 2015, p. 258).

For the start, it is necessary to do a precise definition of the objective of the analysis, because this will have an impact on the later steps. Secondly, the source documents will be selected and uploaded into a working environment to process. It is even more important to structure the data in text analytics because the data needs to be in a format, on which can be worked with the needed techniques. To make this possible, the data needs to be broken down into several words or stems of words. As well as in the source mentioned before, the „bag of words“ is mentioned in this study. Therefore, it is necessary to transform the words into a vector space model representation by using word filtering, lexical processing, syntactic processing, and semantic processing. In a weighted vector shape, it is now possible to use text mining techniques onto the documents to filter the frequency of words (Kinra et al., 2019, p. 177).

The possibility to gain information during text mining in traffic reports will be discussed in the ongoing chapter. The objective is about using historical data from traffic reports to improve the work with google trends. The text mining process in this project is oriented on the text mining process model described in Figure 6: Natural Language Processing (Miller, 2014). Areas of real application will be presented in a shortened version to give some examples.

The project work focusses on text mining in traffic reports. Consequently, it is the database for the text analytics process german traffic reports. The traffic reports are input factors and marked in green color in Figure 7. From the traffic reports the raw text can be extracted due to the big data and workability of the extracts, that must be worked on them. Needed und neglected text processes are further described in this chapter. In the next step are the processed traffic reports classified with the help of keyword analysis. The keyword analysis is more input in the form of a defined text cluster with search words needed. The outcome should be findings of good and bad timeframes on national routes and as a result the possibility to deny transport delays.

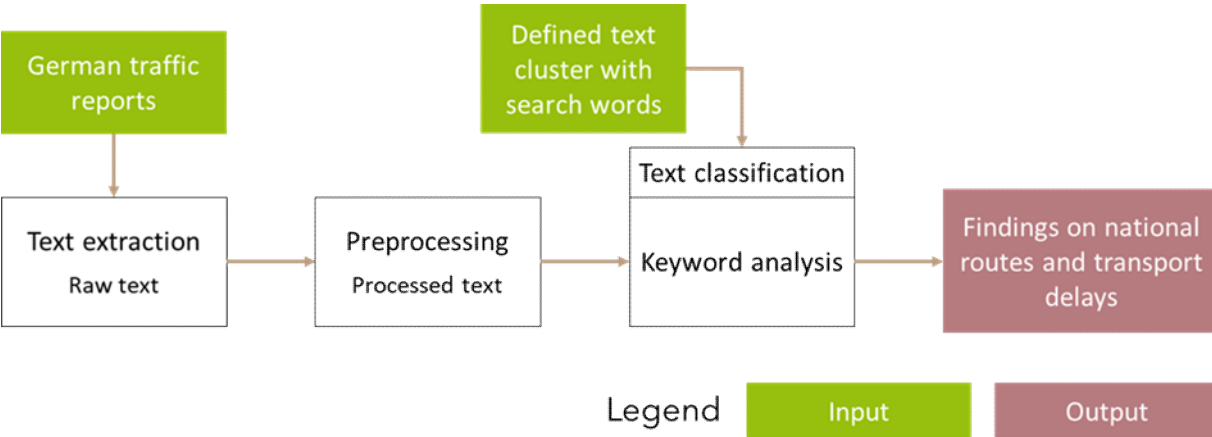


Figure 7: Text mining framework for traffic reports

2.2.7 How has text analytics been used before?

To describe the usage of text analytics before, this paper will focus on two examples from previous researches. That researches topic is, to monitor disease outbreak using the internet or

more specifically, twitter postings by running text analytics over a period of time and compare them to the official influenza monitoring reports from U.S. Centers for Disease Control and Prevention. The second example that will be used is the alcohol sales volume, which is tracked by the U.S. Census Bureau and then compared to, once again twitter postings, but could also relate to other social media platforms (Culotta, 2020)

The data is used to detect influenza outbreaks, comes from the U.S Center for Disease Control and Prevention (CDC) who is publishing weekly reports from the US Outpatient Influenza-like Illness Surveillance Network (ILINet) who combines over 3000 health providers, spread over the whole nation. A big problem with these reports is the suffering from high costs and a slow reporting time. To improve the situation, the research focusses on the text analysis of Twitter messages in combination with other information that is visible on twitter like the city, the state, the gender and the age of the potential patient, which can be very important information when it comes to fighting an influenza pandemic. Analyzing twitter messages in which users describe flu-like symptoms has been shown a strong correlation to the governmental data (Culotta, 2020)

Using Twitter as an alternative source of statistics, the research used keywords related to the consumption of alcohol like "drunk, hangover or hungover" which can be used as potential indicators of drinking alcohol. Because the consumption is not always the same day as the buying of alcohol, a 7-day lag was implemented into the research to monitor a correlation. Those were then compared to the results without the 7-day lag, which led to the summarization, that using twitter postings can be a useful source to explore trends in alcohol consumption (Culotta, 2020)

2.3 Using traffic reports in transport logistics

The text mining framework in this project describes, as shown in Figure 7: Text mining framework for traffic reports, raw text extraction from German traffic reports. For a clear text mining process, it is necessary to know exactly which type of text mining is needed. For the German project, it is quite clear, due to the given HTML source code, which describes the source for the text extraction, that we need web mining.

Web mining is a special type of data mining and can be divided into three different mining processes (Taeho, 2018, p. 13f.):

- Web content mining
- Web structured mining
- Web usage mining

The web content mining refers to mining on the content of web documents. Each web document is classified into its topic as a task of web mining content. The web mining content is the web mining process, which is most familiar with the known text mining process. Another task of

web mining content is to generate a summarization of the web document. Two documents with familiar content are put in the same category (Taeho, 2018, p. 13).

The web structured mining describes a process where structures or templates are considered for the mining task. Two web documents with the same structure, but different content are put in the same category, unlike the web content mining, This means web documents are classified regardless of their content (Taeho, 2018, p. 13).

The third web mining process, web usage mining, describes the web mining where trends of relevant web documents are considered. The web usage mining shows the connection of URL addresses and develops with this information some sort of access behavior of users (Taeho, 2018, p. 14).

For the project work is the web content mining relevant. To get information during web content mining additional techniques of text mining are needed (Taeho, 2018, p. 14). General text mining techniques are explained further and taken into relevance for the study.

Text indexing describes the process of converting an unstructured text into a list of words. This process includes normally three major steps, which are explained below (Taeho, 2018, p. 19f.).

The first step describes a process of segmenting a text into tokens by the whitespace or punctuation marks, the tokenization (Taeho, 2018, p. 21.). Due to the familiar structure of all traffic reports is it possible to combine more tokens into one line of the world list. After the described rule would be “Dreieck Stuhr (58a)”, shown in figure 8, split into three tokens, but its necessary to keep the three words together in one world list line at the end, to find the position on the motorway more easily.

The list of tokens becomes the input for the next steps, stemming and stop word removal.

Stemming is converting the nouns, verbs, and adjectives. Nouns are converted into their regular singular case. The verbs and adjectives are converted into their root forms by removing the postfix (Taeho, 2018, p. 23f).

The third step describes the stop word removal. The process describes only grammatically changes and is irrelevant for the given contents. As stop words can be described as prepositions, conjunctions, and articles (Taeho, 2018, p. 24f).

For the date 01.01.2016 presents the stau1 Archiv 668 traffic reports (“Stau1 Archiv,” 2019.). By calculating with approximately 650 traffic reports a day are $650 \text{ traffic reports} * 365 \text{ days} = 237.250 \text{ traffic reports a year}$. This big number of reports makes it necessary to put the given traffic reports in some sort of normal form.

The first step is to extract the text from the website. For coding means the transformation from an HTML code to a text string. To explain what must be done are the steps presented for one randomly chosen example.



Figure 8: Example of a traffic report for text processing

The extracted text string from Figure 8: Example of a traffic report for text processing looks like this:

„A1 Osnabrück -> Bremen 8:37 Uhr – 23.01.2016 19:17 Uhr Zwischen Dreieck Stuhr (58a) und Bremen Brinkum (57) Gefahr durch Gegenstände 25.12.2015“.

A big advantage in the case of text processing is the fact, that the traffic reports are generated automatically. This gives the traffic reports are defined structure and similar words for defined cases. Due to the similar words for the same cases is no need for text processing in the form of stemming or lemmatization. To reduce the number of words and tokens are the following steps necessary:

- Tokenization
 - Remove punctuation and whitespace
- Remove german stopwords

After the described steps of text processing looks the raw text extract as presented in table 1:

Table 1: Manual extracted word list

Tokens	Explanation
A1	Highway
OsnabrückBremen	Direction
251220150837Uhr	Date and time start
230120161917Uhr	Date and time end
DreieckStuhr58a	Position on motorway
BremenBrinkum57	Position motorway
Gefahr	Issue

Gegenstände	Reason
--------------------	--------

The raw text extract is now a word list. The presented word list includes all the needed information for the classification due to keyword searches.

3 Analysis

3.1 Document selection and structure

To analyze transportation-related problems, its first needed to find accessible data. According to governmental barriers, not every kind of data is open for everybody. This makes it harder to find relevant data to get the right results (Kinra et al., 2019, p. 179).

German traffic reports are saved in a restricted access data archive. The “Stau 1 Archiv” is accessible under this link: <https://stau1.de/stau-archiv/index.php>

The date 01.01.2016 has open access for all users. That’s the reason why given examples are always from the date 01.01.2016 in this term paper. If it becomes real work to implement this framework, there are different abonnements choose able to get access to all the traffic reports. For example costs the business tariff for one month 39.99 € to see all traffic reports in the last 12 months.

The “Stau 1 Archiv” gives different approaches to find the needed traffic reports.

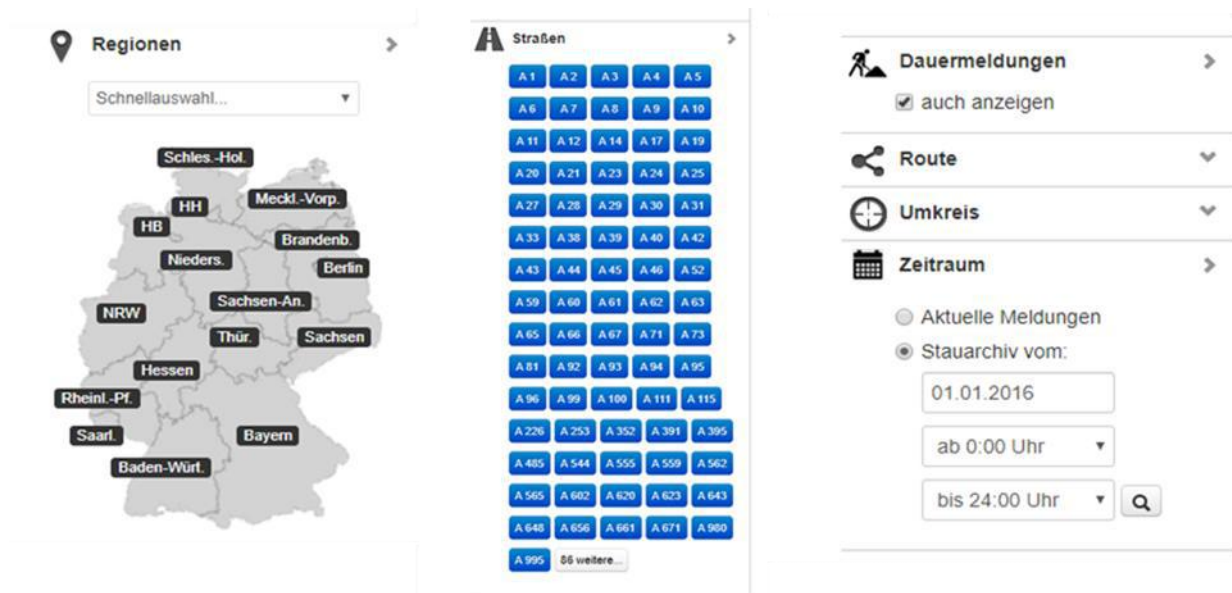


Figure 9: Filter options for traffic reports (Stau1 Archiv)

Figure 9 shows 5 different ways to find the needed traffic reports.

First, you can search on the German map for the relevant regions your traffic reports shall cover. Secondly, you can search for the relevant street names the traffic reports shall present. More ways to search are due to a defined route, so the archive finds on its own where the relevant streets are located and which traffic reports are needed. The same principle works with a defined geographical area.

The fifth way to identify your relevant traffic reports is due to a timeframe you can set. In addition to this its to mention, that you can combine several ways of searching types. It's possible to search with every imaginable combination, e.g. searching for a certain motorway in a defined federal state on a given date.

The number of filter opportunities is important due to the mass of traffic reports collected in this archive. Only for the 01.01.2016, you can find 668 traffic reports for Germany. The mass of traffic reports is manageable even for laypersons due to a lot of opportunities for filtering and similarities in the traffic reports, presented in the next step.

1. Name of roadway and direction
2. Date
3. Exact location
4. In both directions, if that's the case
5. Reason for issue



Figure 10: Structure of traffic reports (Stau1 Archiv)

The headline of every traffic report names the motorway and the direction of the issue on the street, presented under number one in red color in the headline. Secondly, the traffic reports give information about the exact location on the motorway. The issue presented in all three examples is for both directions relevant. In some cases, the issue is only relevant for one direction and then the part in green color is left out. Ending the text frame is marked with blue color and describes the actual situation on the street and mostly the reason for the issue. You can find the timeframe in the header information as well. The light blue painted shows the date and the exact time windows. Sometimes the archive presents persistent information, which lasts for a longer time frame. These traffic reports are listed for every search only one time, even though you search for more than one day. How to work with this type of report is explaining in the Chapter.

3.2 Classification of traffic reports

As mentioned before, the traffic reports are generated automatically, which means for comparable facts are the same words used.

Classification can be divided into two different approaches. The rule-based approach works with manually defined classification rules. The machine learning-based approach where the classification develops by using the sample of data (Taeho, 2018, p. 21)

This study works with the rule-based approach.

Different traffic reports, with different meanings having different importance for the transport of a company. Figure 11 shows how traffic reports are classified in this study.



Figure 11: Classifying of traffic reports

As shown in Figure 11 the traffic reports are split into 4 types of reports. Below is described how to identify these reports due to keyword searches. The marked colors for the ongoing examples can be found in Figure 11 as well. Further is every classification with its own specifications described. The shown colors can be found in the next figures for this chapter as well.

TRAFFIC JAM

The most obvious classification of traffic reports can be clustered in reports describing traffic jams. These reports can be identified by the keyword “Stau”, marked in color red in Figure 11 and Figure 12. The keyword is only used in traffic reports describing traffic jams. In addition to the fact that there are a traffic jam on the street the identified traffic reports presenting in most examples the length of the traffic jam. As discussed before is the structure of the reports not changing. That means that the position of the length number is always in the same position right before the keyword. The defined length could be put in relevance to the expected delay for the company. The longer the traffic jam, the higher is the delay due to traffic jams. This means for the company not every report about a traffic jam has the same relevance for the transport routes. This framework describes only the first level of categorization and neglects further information’s, due to the closed time frame for this project.

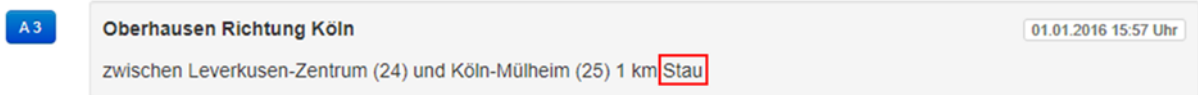


Figure 12: Example of a traffic report for traffic jam (Stau1 Archiv)

SLOW MOVING TRAFFIC

The second cluster of traffic reports describes slow-moving traffic. These reports can be identified by the keyword “stockender Verkehr”, marked in color orange in Figure 13. In comparison to the traffic reports about traffic jams, the reports of this classification presenting the length of the slow-moving traffic as well. The length could be handled the same way the length of traffic jams could be taken into concern. Besides, the length is some reports presenting the issue for the slow-moving traffic. This information is for the first stage of classification neglected as well. The different issues for this case can be found in the annex.



Figure 13: Example of a traffic report for slow-moving traffic (Stau1 Archiv)

DANGER

Thirdly, the most existing case in traffic reports is representing different kinds of danger on the street. Every traffic report using the keyword “Gefahr” is counted as one traffic report in this classification. It’s the most existing case because of the keyword danger, which includes a big number of issues. As shown in the given examples the classification “danger” has no direct relevance for the transportation time, due to no traffic congestions mentioned. But indirectly could the mentioned problems result in slow moving-traffic or in worst case traffic jams. That’s why their classification is still necessary. The relevance of the transport company is not as high as the first two cases but has still to be taken into concern.

The issues shown in Figure 14 are:

- People on the road
- Objects on the road
- Unsecured accident
- Oil track on the road

The different issues are just examples. The whole dataset presents a bigger number of issues that can be found in the annex. The traffic reports are classified in this cluster only by the

information of a certain issue on the roadway, which could result in a delay. Additional information e.g. the mentioned issue is even neglected to the first to cases in this framework.



Figure 14: Example of a traffic report for danger (Stau1 Archiv)

NEGLECTED

Regarding all the traffic reports there are some reports, which are not relevant to the transportation company. These reports can be neglected. Different examples of reports can be found in the annex. These reports are filtered after the principle of exclusion. If the defined keywords of the last three cases cannot be found in the report, the report counts in this category. The given example shows the issue of a closed motorway for heavy load transport with a carriage over 44t. This report is in the category “neglected” because it’s not relevant for typical transportation with a typical weight of the carriage.



Figure 15: Example of a traffic report for neglected reports (Stau1 Archiv)

3.3 Counting of traffic reports

The time in given traffic reports is split into two different categories. On the one hand, it is the time defined for an exact time of the day in the traffic report. This case is easy to count for the defined time. Every traffic reports count the given time of the day. On top of that showing some reports a time frame of a certain number of hours a day. These reports are counted for every hour once.

On the other hand, it includes the traffic report a bigger time window. This case describes permanent traffic issues e.g. constructions, which last for more than a day. This type of traffic reports is mostly classified in the category “danger”. That should be the case of longer terms monitoring. It makes sense to count these reports for every hour in the given window of days if you have monitoring of the traffic reports for a big-time window.

If the time window of filtered traffic reports shorter or only one day the approach must be changed. The reports should be neglected due to the permanence because they don't present any differentiation in time frames if you monitor a short-time window. It would only sum up the cluster "danger" in every hour for the same number.

3.4 Output of framework

The output of the used framework on the described dataset depends on the input and set variables.

As an input must be chosen how the traffic reports are filtered. It depends on the needs of the company. It can be important for a company to have a region or a certain route to monitor. Transport companies mainly know their routes and driven time frames. If that's the case, it's easy for them to generate the right input in the form of traffic reports directly at the website. As described before, it presents the "stau archiv" a lot of different options to filter the needed traffic reports.

The second variable is time. The needed time intervals which interesting or necessary for a company can change.

That's why the time interval can be set individually. Mainly is the more detailed one the better one. The output could be on hourly, half-hourly or even shorter. The counting of the permanent traffic reports would be the same.

For a chosen route and a time frame, the output could look like this. This example describes the framework applied only for Friday the 01.01.2016. This date was chosen, because of its free access. Due to the short-window monitoring are the persistent traffic reports neglected. The relevant traffic reports are filtered about the route option. The starting point is Hamburg and the end of the route is Köln. This filter presents 170 relevant traffic reports for the day. From the 170 traffic reports are 15 neglected due to different reasons. 7 reports are describing persistent issues on the motorway, which last more than one day, as discussed before are they neglected for this case. The framework is only applied to traffic reports for motorways. Reports about federal streets are sometimes presented in the archive as well but not taken into concern.

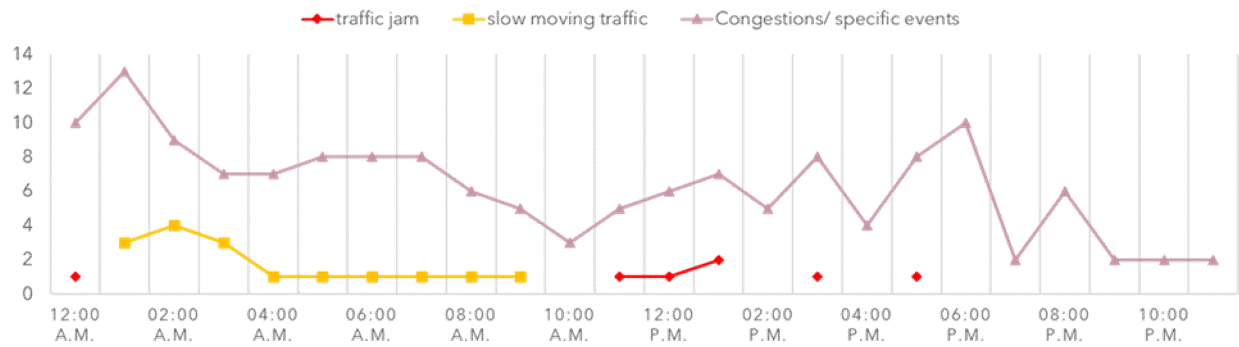


Figure 16: Traffic report 01.01.2016 Route Hamburg – Köln

Figure 16 shows the 170 traffic reports split in the three discussed cluster. Reports about slow-moving traffic are mostly seen in the night and during the morning hours. The reasons for that are on the date 01.01, when winter is in Germany, simply to explain. Fog and poor visibility on the street cause slow-moving traffic mostly. Reports about traffic jams are split over the day with a bigger traffic jam at midday. Congestions, also called Danger in this project, are spread over the whole day with some peaks during the night hours and in the afternoon.

This example shows some trends over the day. If these trends are recurrent for other days in the week or for every Friday of the week cannot be answered in this case. To answer that question must be more dates covered.

4 Conclusion

This project describes a framework to use historic traffic reports for predictive analytics. Actual routing prediction tools, as discussed “google trends” and others, are available. Due to the literature research are the actual prediction tools missing a lot of information to have real reliability for the given prediction. This project presents the potential of gaining one more input factor to increase the reliability of prediction tools. The given framework describes a text mining process to gain historic traffic information about the situation on German highways.

With the findings of the work, the research questions can be answered as followed:

Can google maps be used for transportation problems during the supply chain?

The literature review shows different approaches of companies to use google maps in addition to other tools for traffic prediction. Many examples presenting a loss of usage due to missed information. This gap can be closed a bit more with gained information from historic traffic reports. Referring to the research of Tafidis where it was mentioned that with more data more accurate results could have been achieved using the Google popular Times application, it is getting visible that using text mining could mean possible addition that hasn't been done yet considering transportation issues.

The examples of Tailwind, Routesavvy, and Ubilabs have shown, that the implementation of Google maps data into programs solving the problems within the supply chain has already successfully been taken place. This leads to the conclusion that a further amount of data can lead to even more precise predictions in transport logistics problems.

This leads to the conclusion, that Google maps can be used for smaller predictions, but wasn't designed to solve big transport logistics problems. But having the opportunity to simulate multiple routes at a time with more precise results through more data, a technical implementation becomes conceivable.

How can historic traffic reports be used to predict future traffic?

First of all, it is important to find a good database, where you can extract the raw data from. In Germany, it's possible to gain all traffic reports on one website, which makes it a lot easier.

Secondly, is due to the big data (approx. 237.250 traffic reports a year) a text processing needed. The German reports are generated automatically means they have strict rules for the structure and wording. This has a lot of advantages for text mining and makes some text mining processes not necessary.

In the end, the model is able to find due to keyword research the right classification for the traffic reports. The framework is defined for the German traffic region but a transfer of the framework to other countries could be possible.

In the given work is the framework implemented manual for one example route. The final output shows better and worse time frames for needed routes. Identifying trends need to be more dates covered in the text mining process.

As discussed for the first question, actual prediction tools must develop especially the usage of historical data to generate traffic predictions. The output of the given framework shows the potential to close this gap a bit more and give the prediction tools one more source of input data to calculate with. It is important to mention that the given timeframes are not compatible with traffic prediction without taking other numbers into concern. To improve the prediction tools in the future is a variety of a lot of input factors needed, where the historic traffic reports could present one component.

Weaknesses of the framework are the fact, that there is no differentiation of the timings by the driver on the street. The framework output shows good frameworks for driving the whole route, but the driver is not everywhere on the route at the same time. This means not every report is important for the driver.

Further, it would be the next step for the project to combine the given kilometers for slow-moving traffic and traffic jams with approximately calculated delays for every kilometer. The same procedure could work with traffic congestion where it could be possible to have a probability for resulting in slow-moving traffic or at worst a traffic jam. That could be combined with the calculated delays as well. The output 2.0 would be then in addition to better and worse time frames a calculated delay in minutes for every time frame

References

ADAC, 2018. Staubilanz 2018 - Neue Rekordlängen. URL <https://www.adac.de/der-adac/verein/aktuelles/staubilanz/> (accessed 11.04.19).

Autobahnen in Deutschland - Gesamte Staulänge bis 2019, 2019. . Statista. URL <https://de.statista.com/statistik/daten/studie/200201/umfrage/gesamte-staulaenge-auf-autobahnen-in-deutschland/> (accessed 15.02.20).

Barwig, K., 2013. Entwicklung und betriebswirtschaftliche Bewertung eines Business Engineering-Ansatzes zur Übertragung und Anwendung von Selbststeuerungssystemen in der Transportlogistik, Wertschöpfungsmanagement. PL Academic Research, Peter Lang, Frankfurt am Main.

Culotta, A., 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Lang. Resour. Eval.* 47, 217–238. <https://doi.org/10.1007/s10579-012-9185-0>

Enterprise Transportation Management Software - Products | Tailwind TMS, n.d. Tailwind Transp. Softw. URL <https://www.tailwindtransportationsoftware.com/products/enterprise-tms/> (accessed 15.02.20).

Estrada, M., Roca-Riu, M., 2017. Stakeholder's profitability of carrier-led consolidation strategies in urban goods distribution. *Transp. Res. Part E Logist. Transp. Rev.* 104, 165–188. <https://doi.org/10.1016/j.tre.2017.06.009>

Google Maps – Applications & IT Integration for companies | Ubilabs, n.d. URL <https://ubilabs.net/en/google-maps/integration-for-companies> (accessed 15.02.20).

Güteraufkommen in Deutschland je Verkehrsträger, 2019. Statista. URL <https://de.statista.com/statistik/daten/studie/12240/umfrage/gueteraufkommen-in-deutschland-je-verkehrstraeger/> (accessed 15.02.20).

Heggenberger, R., Mayer, C., 2018. Predictive Analytics in der Mobilitätsbranche, in: Wagner, H., Kabel, S. (Eds.), *Mobilität 4.0 – neue Geschäftsmodelle für Produkt- und Dienstleistungsinnovationen, Schwerpunkt Business Model Innovation*. Springer Fachmedien, Wiesbaden, pp. 1–29. https://doi.org/10.1007/978-3-658-21106-6_1

Kinra, A., Kashi, S.B., Pereira, F.C., Combes, F., Rothengatter, W., 2019. Chapter 8 - Textual Data in Transportation Research: Techniques and Opportunities, in: Antoniou, C., Dimitriou, L., Pereira, F. (Eds.), *Mobility Patterns, Big Data and Transport Analytics*. Elsevier, pp. 173–197. <https://doi.org/10.1016/B978-0-12-812970-8.00008-7>

Kroker, M., 2018. Das gigantische Wachstum von Social Media von 2010 bis 2018. URL <https://blog.wiwo.de/look-at-it/2018/07/25/das-gigantische-wachstum-von-social-media-von-2010-bis-2018/> (accessed 13.11.19).

Mangal, S., n.d. Why You Shouldn't Run a Supply Chain Using Google Maps. URL <https://blog.roambee.com/supply-chain-technology/why-you-shouldnt-run-a-supply-chain-using-google-maps> (accessed 15.02.20).

McKinnon, A., 1999. The Effect of Traffic Congestion on the Efficiency of Logistical Operations. *Int. J. Logist. Res. Appl.* 2, 111–128. <https://doi.org/10.1080/13675569908901576>

Miller, T.W., 2015. *Web and network data science: modeling techniques in predictive analytics* / Thomas W. Miller. Pearson Education, Inc, Upper Saddle River, New Jersey.

Miller, T.W., 2014. *Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R*, Revised and Expanded Edition. FT Press.

Powerful, Affordable Route Planner | RouteSavvy.com, n.d. . RouteSavvy™. URL <https://www.routesavvy.com/> (accessed 15.02.20).

Siegel, E., 2016. Introduction: The Prediction Effect - Predictive Analytics, Revised and Updated. URL <https://learning.oreilly.com/library/view/predictive-analytics-revised/9781119145677/f02.xhtml> (accessed 1.02.20).

Stau1 Archiv, 2019 URL <https://stau1.de/stau/?starttime=2016-01-01%2000:00:00&endtime=2016-01-01%2023:59:59&modal=4> (accessed 21.01.20).

Staustunden auf deutschen Autobahnen, 2020 Statista. URL <https://de.statista.com/statistik/daten/studie/508860/umfrage/staustunden-auf-deutschen-autobahnen/> (accessed 15.02.20).

Taeho, J., 2018. *Text mining: concepts, implementation, and big data challenge*. Springer Science+Business Media, New York, NY.

Tafidis, P., Teixeira, J., Bahmankhah, B., Macedo, E., Guarnaccia, C., Coelho, M.C., Bandeira, J.M., 2018. Can Google Maps Popular Times be an alternative source of information to estimate traffic-related impacts? 97th Transp. Res. Board Annu. Meet.

Taylor, M.A.P., n.d. The City Logistics paradigm for urban freight transport. ResearchGate. URL https://www.researchgate.net/publication/242213289_The_City_Logistics_paradigm_for_urban_freight_transport (accessed 14.02.20).

Transportaufkommen im deutschen Straßenverkehr bis 2022, 2019. Statista. URL <https://de.statista.com/statistik/daten/studie/205940/umfrage/prognose-zum-transportaufkommen-im-strassenverkehr-in-deutschland/> (accessed 15.0.20).

Appendix

Database for Figure 16: Traffic report 01.01.2016 Route Hamburg – Köln

Traffic reports 01.01.2016 (Route Hamburg - Köln)

Neglected: 12 traffic reports

	traffic jam	slow moving traffic	Danger
12:00 a.m.	1		7
01:00 a.m.		3	10
02:00 a.m.		4	11
03:00 a.m.		4	8
04:00 a.m.		1	6
05:00 a.m.	1	1	8
06:00 a.m.		1	10
07:00 a.m.		1	9
08:00 a.m.		1	8
09:00 a.m.		1	4
10:00 a.m.			3
11:00 a.m.	1		7
12:00 p.m.	1		10
01:00 p.m.	2		8
02:00 p.m.			7
03:00 p.m.	1		9
04:00 p.m.			5
05:00 p.m.	1		9
06:00 p.m.			9
07:00 p.m.			6
08:00 p.m.			10
09:00 p.m.			6
10:00 p.m.			2
11:00 p.m.			1

Findings of different keywords from the manual classification of the route Hamburg – Köln

Gefahr	Danger
Keywords German:	Keywords English:
Gegenstände auf der Fahrbahn	Objects on the road
Fahrstreifen gesperrt	Lanes closed
Unfall	Accident
Fahrbahn verengt	roadway constriction
Radfahrer auf der Fahrbahn	Cyclists on the road
Personen auf der Fahrbahn	People on the road
ungesicherte Unfallstelle	unsecured accident site
Gefahr durch Ölspur	Danger from oil spill
durch Blitzeis	by black ice
rückwärtsfahrender Pkw	reversing car
unbeleuchtetes Fahrzeug	unlit vehicle
Tiere auf der Fahrbahn	Animals on the road
Nebelbänke	Fog banks
Personen werfen Gegenstände auf die Fahrbahn	Persons throw objects onto the road
Falschfahrer	Wrong driver
Verschmutzte Fahrbahn	Contaminated road surface
Defektes Fahrzeug	Defective vehicle
Brennendes Fahrzeug	Burning vehicle

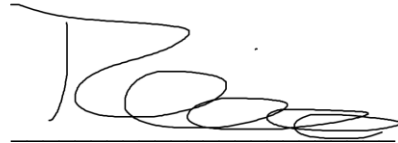
Neglected	Neglected
Keywords German:	Keywords English:
Schwerverkehr	Heavy goods traffic
Dauerbaustelle	Permanent building site
Parkplatz geschlossen	Car park closed
Ausfahrt gesperrt	exit blocked

Affidavit

I hereby declare that I have produced the present work independently and without the use of any aids other than those specified. All passages that have been taken literally or analogously from published or unpublished writings are marked as such. The paper has not yet been submitted in the same form or in extracts in the context of other examinations.

Winsen,

09.06.2021

A handwritten signature in black ink, consisting of a large, stylized 'T' followed by several loops and a horizontal line at the end.

Torben Riemer

Affidavit

I hereby declare that I have produced the present work independently and without the use of any aids other than those specified. All passages that have been taken literally or analogously from published or unpublished writings are marked as such. The paper has not yet been submitted in the same form or in extracts in the context of other examinations.

Bremen,

09.06.2021

A handwritten signature in black ink, featuring a large, stylized 'P' followed by several loops and a horizontal line at the end.

Phillip Warnke