Examining the Potential of Textual Big Data Analytics for Public Policy Decision-making

A Case Study with Driverless Cars in Denmark

Kinra, Aseem; Beheshti-Kashi, Samaneh; Buch, Rasmus; Nielsen, Thomas Alexander Sick; Pereira, Francisco

Document Version
Accepted author manuscript

Published in:
Transport Policy

DOI:
10.1016/j.tranpol.2020.05.026

Publication date:
2020

License
CC BY-NC-ND

Citation for published version (APA):
Kinra, A., Beheshti-Kashi, S., Buch, R., Nielsen, T. A. S., & Pereira, F. (2020). Examining the Potential of Textual Big Data Analytics for Public Policy Decision-making: A Case Study with Driverless Cars in Denmark. *Transport Policy*, *98*, 68-78. https://doi.org/10.1016/j.tranpol.2020.05.026

Examining the potential of textual big data for public policy decision-making on driverless cars: A case study from Denmark

Aseem Kinra, Samaneh Beheshti-Kashi, Rasmus Buch, Thomas Alexander Sick Nielsen, Francisco Pereira

PII: S0967-070X(20)30359-0

DOI: https://doi.org/10.1016/j.tranpol.2020.05.026

Reference: JTRP 2362

To appear in: *Transport Policy*

Received Date: 23 April 2020

Accepted Date: 30 May 2020

Please cite this article as: Kinra, A., Beheshti-Kashi, S., Buch, R., Sick Nielsen, T.A., Pereira, F., Examining the potential of textual big data for public policy decision-making on driverless cars: A case study from Denmark, *Transport Policy* (2020), doi: https://doi.org/10.1016/j.tranpol.2020.05.026.

Examining the potential of textual big data for public policy decision-making on driverless cars: a case study from Denmark

Aseem Kinra[ab1], Samaneh Beheshti-Kashi[a], Rasmus Buch[c]; Thomas Alexander Sick Nielsen[d] and Francisco Pereira[e]

[a]*University of Bremen, Germany, Universität Bremen, Bibliothekstraße 1, 28359 Bremen, Germany*
[b]*Copenhagen Business School, Denmark, Dept. of Operations Management, Copenhagen Business School, Solbjerg Plads 3, DK-2000 Frederiksberg, Denmark*
[c]*ISS A/S, Denmark*
[d]*The Danish Road Directorate, Copenhagen, Denmark*
[e]*Technical University of Denmark, Bygningstorvet, Building 116, room 123A, 2800 Kgs. Lyngby, Denmark*

[1] *University of Bremen, Germany, Fachbereich 07 / Universität Bremen, Global Supply Chain Management, BIBA Gebäude, Raum 1060, Hochschulring 20, 28359 Bremen*
Phone: + 49 (0)421 218-669-60/81
E-mail address: kinra@uni-bremen.de

Examining the potential of textual big data for public policy decision-making on driverless cars: a case study from Denmark

Abstract

The simultaneous growth of textual data and the advancements within Text Analytics enables organisations to exploit this kind of unstructured data, and tap into previously hidden knowledge. However, the utilisation of this valuable resource is still insufficiently unveiled in terms of transport policy decision-making. This research aims to further examine the potential of textual data in transportation through a real-life case study. The case study, framed together with the Danish Road Directorate or *Vejdirektoratet*, was designed to assess public opinion towards the adoption of driverless cars in Denmark. Traditionally, the opinion of the public has often been captured by means of surveys for the problem owner. Our study provides demonstrations in which opinion towards the adoption of driverless cars is examined through the analysis of newspaper articles and tweets using topic modelling, document classification, and sentiment analysis. In this way, the research attends to the collective as well as individualised characteristics of public opinion. The analyses establish that Text Analytics may be used as a complement to surveys, in order to extract additional knowledge which may not be captured through the use of surveys. In this regard, the Danish Road Directorate could find the usefulness while understanding the barriers in the results generated from our study, for supplementing their future data collection strategies.

Keywords*:* Autonomous vehicles (AVs); Driverless cars; Text Analytics; Topic modelling; Big data; Machine Learning; Transport policy

1. Introduction

Big data is increasingly being leveraged within transportation, and under the notion of "*Intelligent Transportation Systems*" the use of sensor technologies, communications (e.g. Wi-Fi, Universal Mobile Telecommunications Systems or UMTS) or GPS tracking lauded as the future of efficient transportation systems. However, considering only this type of structured data as Big Data would result in a huge loss of resources (Kinra et al., 2019). It is widely cited that textual data makes up to 80 per cent of all data being produced, and in the coming years, an exponential growth in the amount of textual content being available for organisations is expected (Fadili and Jouis, 2016). This huge increase of textual data is also fuelled by the emergence of social media platforms such as Twitter (Casas and Delmelle, 2017) or Facebook. Simultaneously, advancements in Text Analytics or text mining enable the exploitation of the previously unattainable textual data sources. In this context, the developments in Text Analytics have opened the door for an alternative big data approach to create information in the transportation sector by exploiting textual data sources. Mining of textual data now plays an important role across politics, health and entertainment, however, in the transportation sector, the application and analysis of textual data is only emerging, limited to the context of social media (e.g. Gal-Tzur et al., 2014), and the extent of its broader potential for transport policy decision making is yet unexplored (Kühl et al., 2019; Jena, 2020).

The objective of this research is to further assess the potential of textual big data for transport policy decision making. We use a mix of different types of the media and employ various Text Analytics techniques to evaluate the potential of textual big data in a real-life case study setting. This involves the examination of the public opinion on the adoption of driverless cars, a current challenge faced by the problem owner, the Danish Road Directorate. The Danish transportation sector faces a range of challenges in the future with more cars on the roads, expected growth in congestion levels and increasing demands for improved mobility, both from companies and the public (Danish Regions,

2017). Additionally, the automation of transportation is also expected to grow, creating new demand patterns as well as opportunities for increased efficiency. Thus, adequately forecasting this future, including the expectations towards and adoption of driverless cars is a key priority for the problem owner. Our analysis and results show that bearing in mind some important identified barriers such as data access and quality, the Danish Road Directorate could find the usefulness in employing unstructured textual big data from newspapers and social media, for supplementing their future (survey-based) data collection strategies.

The rest of the paper is organised as follows. An exhaustive review of the literature is first presented in Section 2, and the main gaps in the literature are identified. Section 3 presents the methodology of Text Analytics and its relationship to content analysis (CA). This is followed by the description of the methodology and the case study, illustrating the main phases and steps which are conducted. The findings are presented in Section 4 and detailed evaluations and implications of the findings are carried out in Section 5. Finally, the paper concludes and offers some future perspectives on the potential of textual big data in transportation.

2. Literature review

2.1. Related works assessing the potential of text in transportation

Only a few works have sought to evaluate the potential of textual data in the context of transportation and no studies have examined this with particular emphasis on Denmark. Rabinovich and Cheon (2011) briefly reflect on the potentials and challenges of CA of textual data, as part of an evaluation of the use of secondary sources in logistics and transportation, but fail to link the use of CA to any specific problems which it can help solve. In Grant-Muller et al. (2015) the potential uses of textual social media data are mapped against generic problems which are faced in the operations of highways and public transport. The study outlines existing methods used for information generation (ANPR camera, RP/SP surveys, etc.) and compares them to uses of social media data. Their paper has a similar purpose as this project and they identify that textual social media data can play a role in the following problems *understanding service quality and driver comfort*, *understanding public opinion* and *detection of undesirable events*. Additionally, they highlight the construction of *origin-destination movements* and understanding *link demand* as areas where textual data may be useful, but where there are significant barriers to overcome first. Unfortunately, they limit their focus on textual data only from social media and only in the context of highway operations and public transportation problems, which means that the usages of multiple other textual data sources and areas of the transport sector are not identified. Their propositions are also not evaluated in applications and the study fails to go beyond identifying possible uses. Subsequently, the barriers or enabling factors for these potentials to materialise have not been outlined. Chaniotakis et al. (2016) also initiated an exploration of the role social media data can play in transportation. They present a SWOT analysis highlighting the potentials and limitations of these sources. Their mapping of the data landscape is based on the capabilities of the platforms and also includes a reflection on the amount of data available and on how this data can be extracted. This mapping is unique in the transportation literature and it serves as a solid point of departure for transportation to discuss social media. However, it is only restricted to social media and unfortunately, their primary focus is on the spatial information (geotags) that can be extracted from these sources, and text is barely discussed. Subsequently, the textual data source landscape is still only vaguely explored for social media sources and traditional sources have not been examined at all.

This shows that the potential of textual data in transportation, in general, has only scarcely been outlined and that there is little existing knowledge to be found in investigating the potential of textual data for the Danish Transportation sector. This, however, does not mean that textual data have not been used through different use cases. In order to understand which problems can be resolved and also to point to where the use of textual data still needs to be explored, we conducted an extensive literature review on the existing studies where Text Analytics approaches have been applied in the context of transportation. This literature review is now outlined and the main findings of the review are discussed.

2.2. An exhaustive review on text mining applications within transportation and the main gaps

In order to ensure an exhaustive review, two specific review methods were used; the database search and the backwards snowballing method (Jalali and Wohlin, 2012). The database search was done in Scopus, which is one of the

main widely-used bibliographical databases. The following query search was done for the keywords/terminologies in the "Title" and "Abstract" of publications index by Scopus to retrieve a reasonable list of publications on transport using CA or Text Analytics:

("Content Analysis" OR "Text mining" OR "Text Analytics") AND ("Logistics" OR "Transport").

In order to reduce the number of publications and increase the relevance to transportation, the search was limited to only include articles in the Social Sciences subject area. This search was supplemented by a search for "content analysis", "Text Analytics" and "text mining" in the most prevalent transportation journals. The journals with the highest SJR ranking in Scimago Journal Rank, which is a publicly available portal that scores the impact of transportation journals[1], was included.

To complement this search, two papers were used for backwards snowballing. Rashidi et al. (2017) and Gal-Tzur et al. (2014) both present literature reviews with a specific focus on the use of social media sources in the context of transportation and they were, thus, suitable as snowballing.

Articles that use CA for classifying existing research on a subject as a part of literature reviews are not included in this review, as they are not directly applied to solve an information problem in the transport sector. For examples of these kinds of publications see Spens and Kovács (2006), Pokharel and Mutha (2009) and Caunhye et al. (2012). Additionally, studies that apply CA as a method to analyse primary data which has been collected in the study are also excluded. For examples of these kinds of articles see Bonet and Paché (2005), Hall et al. (2013) and Combs et al. (2016).

Table 1 presents a consolidated overview of the literature. In general, this review illustrates the variety of potential usages of textual data in the existing literature and that text indeed can be useful and can help to solve a variety of different problems. It is also clear that the emergence of social media and the advancements in Text Analytics methods has expanded potential uses of textual data by solving problems in relation to understanding public opinion, accident causes, and traffic condition evaluation.

However, the review also opened for a number of unanswered questions in the existing knowledge about the potential of textual data in transportation. The first reflection is in relation to the methodological issues in the usage of textual data. Text presents itself as an opportunity to transportation but there is a need for a wider reflection of the barriers to its exploitation which have not been explored in literature. For instance, a range of studies illustrates that textual data, presents an opportunity for capturing user needs and opinions however none of these critically reflect upon the methodological issues and limitations which must be met in order to represent a truthful reflection of users (Collins et al., 2013; Gal-Tzur et al., 2014; Schweitzer, 2014; Wanichayapong et al., 2011).

Second, there is no overview of the sources that contain relevant information, nor how to access this data. A prerequisite for any Text Analytics project is the availability of data and for textual data to be established as a proper alternative for practitioners, there is a need for a mapping of the available sources, how to access these, and which barriers there are in this regard. For instance, regarding the usage of social media data, there are significant privacy and copyright concerns which should be geographically accounted for when evaluating the usefulness of these textual sources, but these barriers are yet to be explored in the transportation literature (Boyd and Crawford, 2012; Lomborg, 2016; Lomborg and Bechmann, 2014).

Finally, the potential applications of textual data within transportation are yet to be subject for a structured evaluation and examination. The review is constituted by studies that apply text as a data source but it is notable how the actual decision-makers are yet to be included in the evaluation of the potential areas in which textual data can help solve problems in the industry. This means its usage is unexplored. Our study takes into account some of these limitations and explores the potential of exploiting textual big data for usage in transportation decision-making problems in the Danish context.

[1] https://www.scimagojr.com/journalrank.php?category=3313

No.	Study	Objective of Study	Case Study	Data Source	Text Analytics Discipline	Algorithm Used	Type of Publication
1	Collins et al. (2013)	Estimating train rider satisfaction	Chicago Train lines	Twitter	Sentiment analysis	Lexicon based classifier (Sentistrenght)	Peer-reviewed journal
2	D'Andrea et al. (2015)	Extracting real time traffic information	Italian road network	Twitter	Document classification	*Support Vector machine*	Peer-reviewed journal
3	Gao & Yu (2016)	Extracting customer satisfaction dimension on public transit	US Public transit agencies	Online reviews	Clustering	LDA topic model	Peer-reviewed journal
4	Gu et al. (2016)	Real-time detection of traffic incidents	Pittsburgh & Philadelphia metropolitan areas	Twitter	Document classification	Naive Bayes classifier	Peer-reviewed journal
5	Pereira et al. (2013)	Accident duration prediction	Expressways in Singapore	Traffic accident reports	Clustering and Prediction	LDA topic model & Regression models	Peer-reviewed journal
6	Schweitzer (2014)	Public opinion on transit operators	American public transportation operators	Twitter	Sentiment analysis	Lexicon based classifier	Peer-reviewed journal
7	Gal-Tzur et al. (2014)	Assessing how transport information from social media can be harvested	Two UK Sporting events in Liverpool	Twitter	Document classification	Support Vector Machine	Peer-reviewed journal
8	Abrahams et al. (2012)	Predicting vehicle returns from car owner forums	Toyota, Honda & Chevrolet forums	Discussion forums	Document classification	Not stated	Peer-reviewed journal
9	Gao & Wu (2013)	Understanding causes of traffic incidents	Missouri state accidents in 2012	Traffic accident reports	Clustering	Verb-based clustering	Conference proceeding
10	Kinra et al. (2016)	National logistics performance appraisal	None	Global supply chain periodical	Document classification	Naive Bayes classifier	Conference proceeding
11	Kosala & Adi (2012)	Extracting real-time traffic information	Jakarta road network	Twitter	Text extraction	Not specified	Conference proceeding
12	Luong & Houston (2015)	Estimating train rider satisfaction	Los Angeles rail transit system	Twitter	Sentiment analysis & Topic modelling	Lexicon based classifier + K-medios	Conference proceeding
13	Maghrebi et al. (2015)	Understanding user mobility	Sydney, Australia	Twitter	Clustering	LDA	Conference proceeding
14	Wang et al. (2015)	Railway equipment condition assessment	Guangzhou railway corporation	Equipment maintenance reports	Clustering & Classification	LDA topic model & Support Vector Machine classifier	Conference proceeding
15	Wanichayapong et al. (2011)	Extracting real-time traffic information	Bangkok urban road network	Twitter	Document classification	Not specified	Conference proceeding
16	Xuan & El-Gohary (2016)	Understanding public opinion on highway projects	Five US highway projects	Public comments on highway projects	Document classification	Several classifiers - HMM, CRF, ME & SVM	Conference proceeding
17	Williams et al. (2016)	Understanding causes of rail accidents	US & Canada	Crash and accident reports	Clustering & Classification	LDA topic model & K-means classifier	Conference proceeding
18	Schulz et al. (2013)	Real-Time Detection of accidents	US cities (Seattle, WA & Memphis)	Twitter	Document classification	Naive Bayes classifier	Lecture notes

Table 1. Overview of Transportation Studies Applying Text Analytics Techniques

3. Methodology

3.1. The Text Analytics methodology

Text Analytics describes techniques for analysing unstructured text data, turning text into numbers to apply statistical models and data mining on a large amount of data, and making the text accessible. Miner et al. (2012) divide Text Analytics into seven practice areas, all of which rely on computerised interactions with text but for different purposes and at different semantic levels. The practise areas are interrelated and a Text Analytics project often relies on technologies from many of these areas. These practise areas are: Search and information retrieval (IR), Web Mining (WM), Information Extraction (IE), Natural Language Processing (NLP), Concept Extraction (CE), Document Classification and Document Clustering. The two latter areas are particularly relevant in the context of creating support for decisions, which is the purpose of this research.

Document classification approaches are divided into machine learning and lexicon-based approaches (Miner et al., 2012). Decision tree classifiers, non-parametric classifiers such as the Support Vector Machine and probabilistic classifiers such as the Naïve Bayes are considered as relevant machine learning techniques for classification tasks (Medhat et al., 2014). An increasing application field of document classification is sentiment analysis and opinion mining (Medhat et al., 2014) which determine the general sentiment of people reflected in the text they are producing and documents are typically scored on a scale illustrating negativity and positivity of a document (Pang and Lee, 2008). Clustering refers to the process of automatically identifying similar items to group them into clusters (Miner et al., 2012). While document classification requires input from both labelled datasets and dictionaries, clustering techniques do not need prior input. One approach of document clustering is topic modelling, which is applied in order to find clusters of co-occurring words in a body of text (Kinra et al., 2019).

3.2. Content analysis and the relation to Text Analytics

CA is the methodology for analysis of text documents which quantify content in term of categories in a systematic and replicable way (Bryman and Bell, 2011). An early definition of this method was provided by Berelson (1952) where CA is referred to as “a research technique for the objective, systematic and quantitative description of the manifest content of communication”. More recent definitions of the technique or method are more inclusive. Krippendorff (2013) defines CA as a research technique for making replicable and valid inferences from text (or other meaningful material) to the context of their use. CA is divided into deductive and inductive approaches. In the first approach, content is split into predefined categories based on a coding scheme, which is based on existing theory. This approach is similar to document classification, in which classifiers obtain a similar role as the human coder in CA. In contrast, within the inductive approach, the categories are developed based on an iterative process where coders go through a portion of the data and then develop the concepts (Elo and Kyngäs, 2008). This process resembles document clustering. The use of computers in CA has been pursued for many years and as early as 1966, first efforts were published in Stone et al. (1966).

The availability of huge textual data sources and the advancements in Text Analytics provide the transportation sector with opportunities to tap into the endless amounts of unstructured data that it generates, and profit from this valuable, but often unexploited data source. CA has been applied to a range of transportation works (Table 2).

Bickerstaff et al. (2002) and Elvy (2014) use CA of policy documents in the UK to evaluate the public participation in a public transport project. Similarly, Lee and Sener (2015) apply CA to a selection of public transportation plans and evaluated the extent to which measures of quality of life are being considered in transport planning. In addition to policy documents, blogs and expert publications have also been used as sources for CA in the transportation area. Casas and Delmelle (2014) analyse a transport-related blog in Colombia to identify potential

sources of transport exclusion of individuals or groups. Kinra (2015) studies a leading supply chain management periodical in order to understand the information measures which are important when conducting country logistics environment assessments. Media publications have also been analysed in relation to a referendum regarding congestion pricing in Canada (Ryley and Gjersoe, 2006). Another transportation-related application of CA worth highlighting is how it has been applied in order to understand the root causes of accidents. Among others, Newnam et al. (2017) analyse coronial inquests and traffic reports from road accidents in Australia.

Table 2. Transportation studies applying CA

Authors	**Study objective**	**Applied sources**	**Geographical Focus**
Bickerstaff et al. (2002)	Evaluation of public transportation in public transportation projects	Policy documents	UK
Elvy (2014)	Evaluation of public participation in public transportation projects	Policy documents	UK
Lee and Sener (2015)	Evaluation of the extent to which measures of quality of life is being considered in transport planning	Public transportation plans	USA
Casas and Delmelle (2014)	Identification of potential sources of transport exclusion of individuals or groups	Transport related Blog	Colombia
Kinra (2015)	Country logistics assessments.	Supply chain management periodical	Global
Newnam et al. (2017)	Identification of root causes of accidents	Coronial inquest/traffic reports	Australia

Therefore with the increased volume of textual data, text mining as the equivalent of an automatised CA may enable access to untapped information assets (Kinra et al., 2016). The research seeks to evaluate and demonstrate whether text analysis of tweets and newspaper articles can be used as a supplement to measure public opinion, in addition to traditional methods such as surveys to measure the general public concerns and expectations towards driverless cars.

3.3. Action research and real-life case study

An interventionist design akin to the action research approach that often involves real-life problems and case studies (Coughlan and Coghlan, 2002) in the area of logistics (see also Näslund, 2002 and Näslund et al., 2010), was adopted as the main method. Following this a methodology consisting of three phases has been developed (Figure 1).

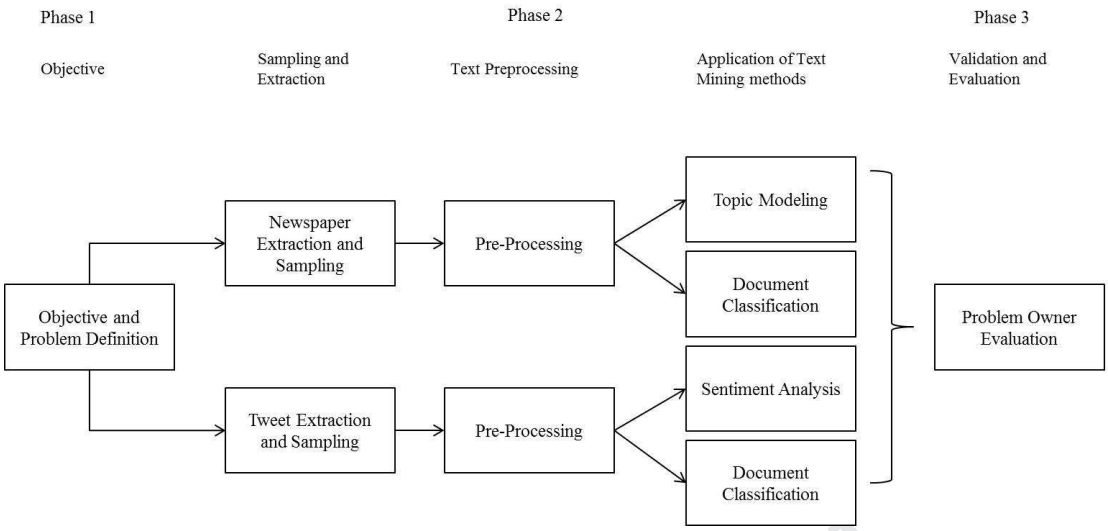

Fig. 1. Methodology Case Study Approach**.**

In the first phase, the objective and scope of the analysis are defined. The second phase consists of the data extraction and sampling, the pre-processing of the generated datasets and the application of text mining methods on two sources. In the last phase, the findings are evaluated and validated within a feedback group session with five experts on driverless cars at the Danish Road Directorate.

Phase 1: Objective and Problem Definition

The problem, which was framed together with the chief consultant from the Road Directorate as the problem owner, was to gain a better understanding about the public's opinion about the benefits and barriers in the adoption of driverless cars, through the employment of Text Analytics. Newspaper articles and Twitter have been selected as two relevant sources of examination. As a first step, a critical literature review was conducted to examine the existing knowledge about the adoption of driverless cars and to have a reference which can be used to see if textual data and Text Analytics are suitable to generate new knowledge for the problem owners (Kühl et al., 2019).

Phase 2: Text Analytics on Newspaper articles and Tweets

Examining the public opinion on driverless cars in the Danish newspaper coverage

In order to extract the main topics discussed in the context of driverless cars in the Danish media, topic modelling is selected as a suitable Text Analytics method. For this purpose, a newspaper corpus was generated. The media articles were extracted using a Danish media database (Infomedia) and in total 1,338 publications were analysed. The search term used to extract these articles was "selvkørende bil*" (self-driving car) or "førerløs bil*" (driverless car) and to improve the accuracy of the search, a filter was applied to only include the articles which mention the keyword in either the headline or the outline of the articles. The articles were extracted over the last five-year period (7th of February 2012 – 7th February 2017). All Danish media publishers both print and web media are included and all media types are included except for Radio and TV and news agencies (Reuters & AP). The same articles are frequently published in different newspapers who have the same owner and in this extraction, each unique article was only included once.

The text pre-processing has been conducted using the Konstanz Information Miner or KNIME analytics platform

version 3.3.1, following similar steps employed in prior studies (Guo et al., 2017; Pereira et al., 2013; Tirunillai and Tellis, 2014). Subsequently, the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) algorithm has been employed to extract topics from the articles.

The output of LDA is a selection of topics which are described by keywords that occur in the topic. The naming of the clustered topics was first conducted by one researcher and then confirmed by a second researcher who was not otherwise related to the project. Naming was based on the identification of a logical connection between the most frequent words for a topic. Following this approach, the naming of the overall clusters is conducted in an exogenous way. However, it is notable, that this applies only to the naming of the clusters. The generation of the topic models is an endogenous process. The topics were visualised using the Fruchterman-Reingold algorithm which is a force-directed layout algorithm that presents the topics as a network (Fruchterman and Reingold, 1991).

In addition to the topic modelling, a lexicon-based document classification has been conducted on the newspaper dataset, in order to track the effects of driverless cars described in the literature. This step is considered relevant in terms of evaluating the potential of Text Analytics in generating additional knowledge, compared to existing methods.

Examining the public opinion on Driverless Cars on Twitter

The second sample analysed was extracted from the microblogging service Twitter. The sample consists of 157,000 tweets which were collected in a 5-month period between the December 2016 and April 2017 using the official Twitter Automatic Programming Interface (API). We searched for Tweets mentioning the terms “Driverless cars”, “Autonomous vehicles” or “Self-driving cars”. The pre-processing of the tweets is carried out using KNIME for removing numbers, converting cases, erasing punctuation and removing common stop words.

In contrast to the newspaper data set, further pre-processing was required, since social media data typically contains widespread use of abbreviations, emoticons, and misspellings (Kumar et al., 2014). Subsequently we applied sentiment analysis in order to investigate how the public was tweeting about different themes in the context of driverless cars. This was done making use of the SentiStrength software, and sentiment lexicons were used (http://sentistrength.wlv.ac.uk; Thelwall et al., 2010). In addition, document classification on the tweets has been conducted, following the same dictionary-based approach as for the newspapers.

Phase 3: Evaluation and validation

The last phase of the case study was dedicated to the evaluation and validation of the findings. A feedback session with 5 experts from the Road Directorate was conducted for this purpose where the experts were presented with findings from the analysis. The purpose of this step is two-fold, as first it served for validation purposes where the session was designed in order to verify the face validity and provide validity for the findings of the analysis. This step is also consistent with the requirements outlined in the CA literature (Krippendorff 2013; Riffe et al. 2005). Secondly, the session sought to provide a critical evaluation of the limitations and methodological barriers of using Twitter and newspapers as sources of public opinion.

4. Analysis and findings

4.1. Public opinion on driverless cars captured through existing surveys

To enhance knowledge about the general public perception of driverless cars, it is necessary to analyse the ways the existing body of knowledge about public opinion on driverless cars has been constituted. Public attitude and

opinions about driverless cars have commonly been examined through traditional survey methods. As the research was conducted in 2017, the most important studies in the context of the research until that year are briefly outlined as follows.

Schoettle and Sivak (2014) surveyed approximately 1,500 persons in the US, UK and Australia and more than 55 per cent of respondents affirmed their positive general attitude to driverless cars. However, a majority also expressed general concerns towards riding driverless cars with only 12 per cent stating no concerns at all. A more positive public attitude has been found in a primarily Austrian survey conducted by König and Neumayr (2017). In the Danish context specifically, the attitude towards driverless cars has been examined by the Danish Road Directorate in a survey with 3,000 Danish respondents in 2017. This survey illustrates a more hesitant attitude towards driverless cars. Table 3 provides an overview of the most relevant surveys about public attitude to driverless cars. By comparing the findings of the studies, all of which examine benefits and concerns, it is evident that four specific concerns are prevalent among the public across the different geographies and studies.

Safety in terms of equipment failure and technological capabilities, the legal liability in case of accidents, the risk of vehicles being hacked and data privacy are found in all of the five studies as specific concerns across the populations. In addition to these concerns, the impact on the labour market and risk of misuse for terrorism have only been highlighted in a single study each and this indicates that these are not widespread concerns according to the surveys. One of the issues in the production of knowledge through surveys is that the researcher defines pre-set questions and categories based on their own understanding of the problem. The concerns and benefits in the studies have therefore all been found through predefined questionnaires, and subsequently, concerns and benefits which were not asked for might be significant but not explored. Furthermore, it is also unclear how important the different concerns and benefits actually are and which the most pressing of the concerns and benefits cannot be derived.

Table 3. Overview of surveys about public attitude to driverless cars

Study	Bansal and Kockelman (2016)	Schoettle and Sivak (2014)	Vejdirektoratet and Wilke (2017)	Kyriakidis et al. (2015)	König and Neumayr (2017)
Response demographics					
Number of respondents	1,088	1,533	3,040	5,000	489
Geography	Texas, US	UK, US & Australia	Denmark	Global	Austria
Other comment on demographics	Representative of state	500 from each country	Representative of nation	From survey platform	66 per cent < 30 years
Benefits expected from Driverless cars					
Improved Safety	✓	✓	✓		✓
Lower Congestion	✓	✓			✓
Lower Environmental impact	✓	✓			✓
Use time for secondary tasks			✓	✓	✓
Improve mobility of weak drivers					✓
Concerns over Driverless cars					
Equipment failure	✓	✓	✓	✓	✓
Legal liability	✓	✓	✓	✓	✓
Hacking	✓	✓	✓	✓	✓
Data privacy	✓	✓		✓	✓

Labour market		✓
Misuse (Terror)	✓	

4.2. Text analysis findings 1: Topic Modelling of Newspapers

The first analysis involved a topic model clustering of newspaper articles. Topic modelling as an unsupervised clustering algorithm enables exploring concerns of and benefits to the public, without the bias of a prior input. The outcome of the LDA is illustrated in Figure 2.

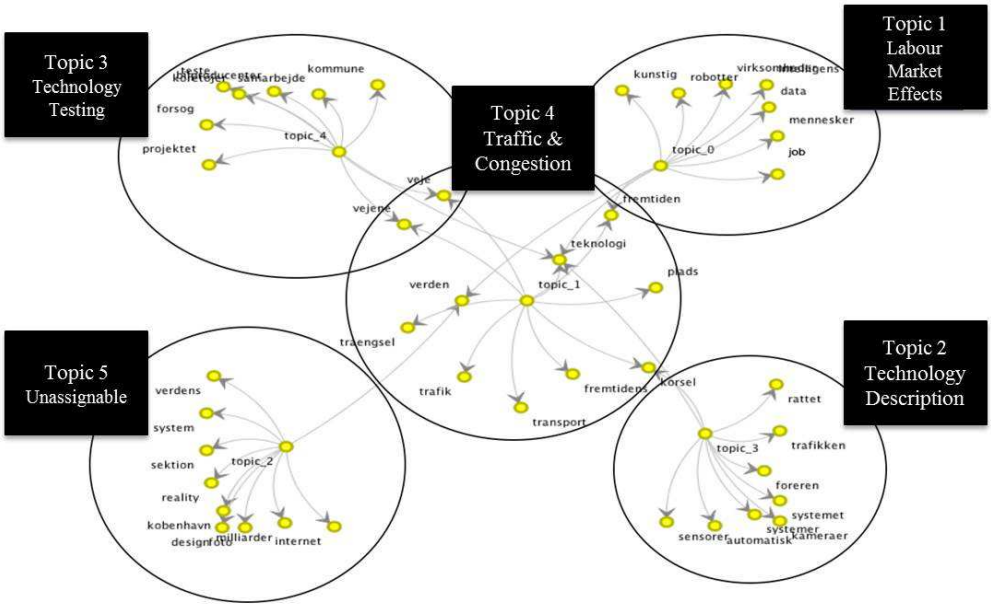

Fig. 2. Clusters of topics in Danish Newspapers coverage of driverless cars

Both topics 1 and 4 *i.e.* labour market effects and traffic and congestion may be considered as more relevant as opposed to topics 2 and 3, because these directly convey the effects and impacts of driverless cars. The occurrence of topics 1 and 4 indicates that there is an interest in how the adoption of driverless cars will influence jobs and the labour market, and whether or not the adoption is related to (more or less) congestion.

It is not possible to reflect on the themes that did not occur, as the output of the topic model is influenced by the similarity in which a given topic is described and subsequently this uncertainty does not allow for conclusions about themes that are not present (Blei et al., 2003). However, it is notable when comparing it to the existing surveys that labour market effects, which are only found to be of public interest in one of the studies outlined, stand out as a topic. This suggests that the effect driverless cars will have on the labour market is more important to the public than what has been shown in existing surveys.

4.3. Text analysis findings 2: Document Classification of Newspapers

The second analysis of the newspapers has been document classification in which the paragraphs of the papers

are categorised into the categories of benefits and concerns found in the surveys. This enabled a ranking of the topics based on the amount of coverage. The analysis shows that safety (424) is the most important theme with almost double the amount of coverage as second and third most covered themes, congestion (236) and labour market (188). The question of liability (120), as well as environmental impact (119), were found to be of significant interest to the public. Hacking (35) of vehicles and data privacy (27) issues were found to be the least important themes as they barely appeared in any paragraphs.

Comparing these results with the survey results, it is clear that safety stands out as the most important theme because safety-related effects of driverless cars consistently emerged as both benefits and concerns across all surveys. However, this analysis indicates that the concerns about hacking and data privacy which though found in almost all existing surveys are not all that important when measured by the term's appearance in the media. In this context, it is noteworthy that while labour market effects are being discussed in three times as much intensity (as measured by the number of paragraphs) as Hacking and Safety combined, in the surveys, these were only found to be important in one out of five studies. This supports the findings from the topic modelling and further indicates that labour market effects may be more important to the public than what has previously been found by means of surveys.

4.4. Text analysis findings 3: Document Classification and Sentiment Analysis of Tweets

Next, document classification and sentiment analysis were conducted on the tweets. Figure 3 illustrates the results from the document classification and sentiment analysis, which offered an in-depth assessment of the themes as volumes indicate importance and sentiment scores indicate the extent to which a theme is perceived as being concerning or beneficial.

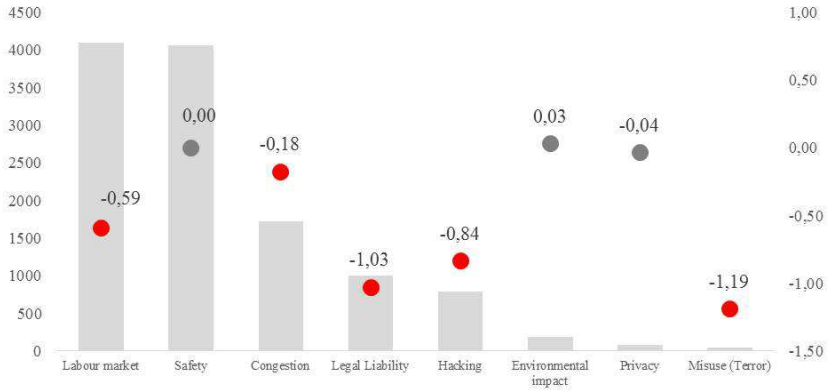

Fig. 3. Number of Tweets classified and average sentiment per topic

Comparing the importance derived from the tweets with those of the newspapers in Denmark, the four most important categories (congestion, labour market, legal liability, safety) are found to be the same which underpins the reliability of the findings. In order to assess the sentiments of the tweets, it is suggested to conduct the sentiment analysis on the level of the topics, and not on the aggregated level of the tweets, since on the aggregated level the scores do not add many benefits to the analysis. Though, by tracking the scores for the classified tweets it is possible to derive the sentiment of the public for specific themes, which are indicated with points in Figure 3.

The analysis of the Danish newspapers shows that Safety and Congestion are the two most important topics but additionally Labour Market effects also stand out as an important topic. The analysis of the tweets illustrates a

similar picture with Safety, Labour Market and congestion as the most important topics. In contrast to the existing literature on public opinion about driverless cars, the analysis of the Tweets and the Newspapers illustrate all a somewhat different picture of the public attitude towards driverless cars. The two findings which especially stand out in this regard are first, that throughout the analyses Labour Market Effects are found to be a significant concern for the public, but it is only found to be a concern in the survey conducted by König and Neumayr (2017). Additionally, throughout the surveys concerns about Hacking, Legal Liability and Privacy issues are shown to be highly important but consistently the analysis of secondary textual data (newspaper and Tweets) indicate that these concerns are less important than initially thought. One explanation why Hacking, Legal Liability and Privacy issues did not emerge through the topic modelling approach may be related to the limitation of five topic models. Increasing the number of allowed topic models may result in new topics in this direction. The analysis is consistent with the surveys in relation to Safety as it also finds it to be a highly relevant topic for the public and additionally illustrates how the public is both concerned about the prospect of autonomous vehicles but also expects them to have the general positive effect of traffic safety.

5. Evaluation of results, discussion and limitations

In order to evaluate the extent to which these types of analyses produce an accurate representation of the public, the possibilities and limitations must be scrutinised in more detail. Accordingly, a feedback session with five experts on driverless cars was conducted. While the experts have been positive about the main findings from the analysis, they have also shown concerns regarding the following issues for the usage of newspapers: over-representation of niche media and a general need for references/experiences on the relevance and representation of the "media population" towards the purpose of representing popular opinion, effects of lobbying, the impact of newsworthy events, and the access of newspaper content since it requires intense resources. As barriers to using Twitter data for accessing public opinion, the experts pointed out the insufficient Danish data sample and the inability to determine sample composition. Similar to the newspapers, the sensitivity of the timing of the sample was a concern, and brings up the question of motivations and drivers behind tweeting behaviour. Table 4 illustrates the main concerns and how these can be tackled, and these are now discussed in more detail in the following sub-sections. The identified barriers are then complemented by the methodological issues raised in the literature, not only limited to the field of transportation. In this regard, methodological concerns are often formulated related to sample validity and sample bias of online textual data (Webb and Wang, 2013; Kim et al., 2013; Gerlitz and Rieder, 2013; Morstatter et al., 2013a; Morstatter et al., 2013b; Tufekci, 2014; Cihon and Yasseri, 2016; Peffer et al. 2018) but also in general related to the use of text analysis (Zanini and Dhawan, 2015). Furthermore, concerns about credibility, privacy, and security issues are reported (Webb and Wang, 2013; Kumar et al. 2016; Costantino et al., 2017; Hassan, 2018). Similarly, Kobayashi et al. (2018) advocate the establishment of reliability, validity, and credibility of output generated through the application of text analysis in the context of organisational research.

Table 4. Main Methodological Issues and Barriers in the Text Analysis of Newspapers and Twitter

Barriers/Methodological issue	**Example of an issue from the case study**	**Is the issue avoidable or controllable?**
Newspaper		
Sampling Bias	Overrepresentation of niche media	Yes, either by including only nationwide papers or weighing the importance based on circulation figures

Content validity	Volumes are highly influenced by newsworthy events and interest in the topics may vary over time	To some degree, by manually controlling for impactful events
Content validity	Lobby effects influence what media publish	No, lobby impact is not controllable
Twitter		
Sampling validity	Inability to determine sample composition in Twitter data	No, Twitter does not provide necessary information about users
Content validity	Volumes are highly influenced by newsworthy events	To some degree, by manually controlling for impactful events
Data availability	Insufficient amount of Danish data on driverless cars	Yes, this issue is determined by the contextual factors topic and geography

5.1. Sampling Bias in the Newspaper Analysis – An Over Representation of Niche Media

The first issue raised was related to sampling bias. In the analysis, all articles independently of the publisher, are assigned the same level of importance. This means that no matter if a paragraph is posted in a nationwide newspaper or in a niche magazine, they are weighed with the same importance. The basic assumption behind applying for newspaper coverage as a measure of public opinion is that the newspapers, on the one hand, write about what concerns the public and on the other hand influence what the public is concerned about, thus, establishing causality. In the case of driverless cars, motor and technology-focused magazines will report a lot about driverless cars and, thus, be overrepresented in the sample, given that they are only read by a specific subsection of the population. This means that the sample will be biased towards the opinion and interest of the most frequently reporting publishers even though these do not represent all of the population which the analysis seeks to understand. Even though this is a definite issue in the analysis in the case study, it is possible to mitigate this bias in two ways: a) either by only including nationwide newspapers or b) by weighing the importance of the newspapers in the analysis based on their circulation figures.

5.2. Content Validity of the Newspaper Analysis – The Effects of Lobbying

Another issue that was highlighted was that individuals or organisations may be able to push forward a specific agenda or theme, and this might be overrepresented. This point was exemplified by the expert in terms of how his organisation has been pushing the congestion theme, which may prove why the theme stood out both as a topic in the cluster analysis as well as the second most frequent category in the classification analysis. Content validity is the degree to which a measure demonstrates the behaviour or phenomenon for which it is intended (Belderbos et al., 2017) and if lobbyism influences the number of newspaper articles which are published on a given subject, then measuring the number of publications to determine public attitude might not be as valid as originally assumed. One measure which can indicate the extent of such influence is how frequently specific organisations and persons are cited, and in the sample applied in this study, the Road Directorate is cited in 68 out of the 1,400 articles. However, even when the number of citations can be tracked the motivation or reflections behind what journalists choose to publish may not be apparent. This means that it is not possible to control for this effect and subsequently, this is found to be a general limitation to the use of newspaper data as a reflection of public attitude.

5.3. Content Validity of the Newspaper Analysis – How Different Themes Develop Over Time

Another content validity related issue raised by the experts was that of aggregation of time period of the utilised sample. Three out of the five experts, therefore, highlighted the need to control the development of themes over time, e.g. whether the coverage of a particular theme has happened in a short space of time due to newsworthy events or if the interest has been more constant throughout the sample. Apropos, there might be a connection between the themes and the level of maturity of the technology. Subsequently, the importance of the driverless car concerns such as hacking and data privacy may be low due to fact that technology is quite far from being a real option for the public and, thus, these kinds of issues are not yet found to be important, but might be at a later stage. Regarding this limitation, text analysis does indeed provide the possibility to track the development of themes over time and the possibility to control for spurious effects (Kinra et al., 2019).

5.4. Sampling Validity of Twitter Data - Inability to Determine Sample Composition

Another critical issue raised by the experts was related to the representativeness of the users in the sample. One of the principle foundations of any valid CA is that the sample is representative of the population and sampling validity refers to the degree to which the collection of data contains, with a minimum of bias, a maximum of relevant information about the universe, correcting particularly for the bias in their selective availability (Krippendorff, 2013).

However, it has been pointed out that Twitter users are younger, better educated and comprise of overrepresentation of people living in cities (Culotta et al., 2015; Mislove et al., 2011). Moreover, only a fraction might be actively expressing because it has also been pointed out that about 40 per cent of active users sign in just to listen and read tweets (Twitter, 2011). Additionally, individuals may make use of several accounts, newspaper agencies are sharing their news stories and companies are also expressing opinions on Twitter, and lastly, some accounts may be robots who are programmed to retweet and write certain messages. In fact, it has been estimated that between 9 to 15 per cent of all users on Twitter are robots (Varol et al., 2017).

In this specific example, the sample consists of 75,000 individual users and these issues would not be a problem if it was possible to constitute a sample that controls for these biases and, thus, reflects the average population. However, Twitter is strict towards sharing personal information about their users and the official API only returns the username of a twitter account, and not the real name, nor any personal information. This means that assessing whether a user is actually a member of the public, a news agency or in fact a robot is not possible by other means than going through each user's self-provided description on twitter. With a sample with 75,000 different users, this remains a cumbersome option. Though the other demographic information such as age, occupational and educational background as well as other relevant sampling criteria would still be unattainable, as this information is seldom shared by users on Twitter.

Another issue related to the sample composition is that the same user may post very frequently about the topic and, thus, be included multiple times in the sample and, thus, have more weight in the analysis. In the case study the most frequent user, TLWNewsPump, contributed with 1,837 tweets in the sample, which accounts for more than 1 per cent of the total sample. Whereas about 11 per cent of the most active users accounted for approximately 50 per cent of the entire sample. It is possible to control for this when doing sentiment analysis by averaging the sentiment per user, however, in classification analysis the results are binary and this effect cannot be controlled for. To mitigate and minimise this issue, one could remove all users who have more than a certain number of tweets, however, this would imply that users are disqualified merely because they are active in a given topic, which could also contribute to other methodological issues.

The sampling barrier brought out by the experts, is also visible in the corresponding literature. Indeed, the representativeness of Twitter data is a widely discussed barrier for its use. Cihon and Yasseri (2016) even point out that the API acts as a "black box" as it does not provide representative data. This also goes along with further researchers' estimations in terms of the limitation and the incapacity the API to deliver scientifically sound random samples as advocated also in early research by Morstatter et al. (2013a), Morstatter et al. (2013b) or in a recent study by Pfeffer et al. (2018).

5.5. Content Validity of Twitter Data - What Triggers Tweeting Behaviour?

The third issue which was pointed out by the experts was similar to the one with newspapers, content validity due to the sensitivity of the timing of the sample. What drives tweeting behaviour? Taking the tweets classified in the safety topic as an example, it is evident that external events had a strong influence on the number of tweets in our sample. Notably, three events generated spikes in our sample, and collectively generated 1,177 tweets which were more than a fourth of all the tweets about safety. These three events constituted a safety stunt by the Chinese technology giant Baidu, a driverless roborace car crash, and an Uber driverless car crash. Similar patterns could be seen in the volumes of the other categories too.

An implication of this is that when evaluating tweet volumes, as in document classification analysis, one should be aware of the influence of stand-alone events. Similarly, it can also be questioned whether the sheer volume of tweets about a topic is a good measure for the actual importance of a topic, or rather for an event that has taken place.

Consulting the literature in this regard, the selection of adequate hashtags or search keywords plays an important role as this selection can have an impact on the volume of the retrieved content (Kim et al. 2013). It may happen that the selected search terms are not or only fairly related to the targeted topic as in Bosley et al. (2013). Kim et al. (2013) suggest a qualitative review of the retrieved content to prevent such effects and to ensure that the selected terms return relevant content. Focussing only on hashtags for creating a dataset can result into biases as discussions for instance on events, can be continued on Twitter without the use of specific hashtags referring to that particular event (Tufekci, 2014).

5.6. Data Availability on Twitter - Insufficient Danish Data Samples

As in the case of newspapers, some distinct methodological issues regarding the use of Twitter were also raised by the experts. First, the volume of the found tweets was too low, only around 50 tweets were returned weekly. Second, the users publishing tweets about driverless cars in Denmark constituted primarily of politicians, representatives of user associations, journalists, news outlets or stakeholders of the transport sector. This means that a specific barrier to using tweets as a measure of the Danish public opinion is the lack of available data and therefore a global selection of tweets in English had to be used for the analysis. When extracting tweets, it is possible to determine geographical boundaries and subsequently, it is possible to limit the perspective from global to countries who are assumed to be contextually similar to Denmark. However, without being able to establish a Danish baseline, it is not possible to understand the influence of these geographical discrepancies. The low volume of Danish Tweets and subsequent inability to gather a Danish sample is found to be a general limitation to the use of the Twitter data in Denmark, though countries like England, Canada, and the US may not face similar issues. Besides the geographical determination, one way to get an overview on the availability of data prior to the generation of the dataset is to define the hashtags or search keywords carefully.

6. Conclusion and future perspectives

The objective of this research was to assess the potential of textual data in transportation. For this purpose, a case study employing Text Analytics for the extraction of public opinion on driverless cars was conducted. Our analysis demonstrates three ways in which Text Analytics can be used to generate knowledge about public opinion. First, topic modelling can be applied to identify the key topics of a text corpus. Second, document classification can be used to categorise the documents based on which topic they relate to. Third, sentiment analysis can be applied to assess the affective state of a text and, thus, evaluate how the sender feels about the topic which they are discussing.

The analysis of the Danish newspapers illustrates that safety and congestion are the two most important topics, though labour market effects has also emerged as an important theme. The analysis of the tweets illustrates a similar picture with safety, labour market and congestion as the most important topics (cf. Fagnant and Kockelman, 2015). In contrast to the existing surveys on public opinion about driverless cars the analysis of the Tweets and the newspaper illustrate all a somewhat different picture of the public attitude towards driverless cars. The major finding which emerged in this regard is that throughout our analysis we found labour market effects to be a significant concern for the public. This is in contrast to the previous findings where only one prior survey, conducted by König and Neumayr (2017), identifies these effects as a concern.

In order to ensure the validity of the results, several measures have been applied. Regarding the application of Text Analytics, two different text sources have been selected. In addition, several text analytic techniques have been applied. Moreover, the problem owner was included in framing the research objective, and evaluating the results in a feedback session with experts in the field and the barriers and limitations have been discussed and presented. Consequently, in the existing case study the usage of textual data, has resulted in the extraction of additional knowledge on the adoption of driverless cars from the public. According to the experts, this additional information can be used by the Danish Road Directorate for evaluating their strategies on the adoption of driverless car more precisely.

Looking forward, while the techniques employed in this study have been used before in transportation e.g. topic modelling for demand prediction in special events (Rodrigues et al., 2017), incident impact prediction (Pereira et al., 2013) or for detecting urban functional zones (Yuan et al., 2012), the field is still in its infancy. A particularly under-explored, yet very promising, opportunity is that for complementing survey analysis, and this paper provides a case study in this direction. Furthermore from a methodological perspective, a new set of techniques, from the deep neural networks (DNN) field deserves attention. An increasingly popular tool in DNN is the “text embeddings”, which provides a cloud of related words, given an input word and context (e.g. Stanford GloVe (Pennington et al., 2014)). This provides associative connections between words and a richer way to represent text.

Acknowledgements

Earlier versions of this research have gained from presentation at the Bremen International Conference on Dynamics in Logistics (LDIC) 2018, and at the WCTR 2019. We thank the Experts at the Danish Road Directorate for participating in the study, and also acknowledge the reviewers and the SI editors for their helpful feedback.

References

Abrahams, A.S., Jiao, J., Wang, G.A., Fan, W., 2012. Vehicle defect discovery from social media. Decision Support Systems 54 (1), 87–97. https://doi.org/10.1016/j.dss.2012.04.005

Bansal, P., Kockelman, K.M., 2016. Are we ready to embrace connected and self-driving vehicles? A case study of Texans. Transportation 45, 641–675. https://doi.org/10.1007/s11116-016-9745-z

Belderbos, R., Grabowska, M., Leten, B., Kelchtermans, S., Ugur, N., 2017. On the Use of Computer-Aided Text Analysis in International Business Research. Global Strategy Journal 7, 312–331. https://doi.org/10.1002/gsj

Berelson, B., 1952. Content Analysis in Communication Research. Pp. 220. Glencoe, Ill.: The Free Press. The ANNALS of the American Academy of Political and Social Science 283 (1), 197-198. https://doi.org/10.1177/000271625228300135

Bickerstaff, K., Tolley, R., Walker, G., 2002. Transport planning and participation: The rhetoric and realities of public involvement. Journal of Transport Geography 10 (1), 61–73. https://doi.org/10.1016/S0966-6923(01)00027-8

Blei, D.M., Ng, A.Y.A.Y., Jordan, M.I.M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3 (4–5), 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Bonet, D., Paché, G., 2005. A new approach for understanding hindrances to collaborative practices in the logistics channel. International Journal of Retail & Distribution Management 33 (8), 583–596. https://doi.org/10.1108/09590550510608386

Bosley, J.C., Zhao, N.W., Hill, S., Shofer, F.S., Asch, D.A., Becker, L.B., Merchant, R.M., 2013. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation 84 (2), 206–212. doi:10.1016/j.resuscitation.2012.10.017

Boyd, D., Crawford, K., 2012. Critical Questions for Big Data. Information, Communication & Society 15 (5), 37–41. https://doi.org/10.1080/1369118X.2012.678878

Bryman, A., Bell, E., 2011. Business research methods (3ed ed.). Oxford University Press, oxford, United Kingdom https://doi.org/10.1016/B978-0-12-387000-1.01001-9

Casas, I., Delmelle, E.C., 2014. Identifying dimensions of exclusion from a BRT system in a developing country: a content analysis approach. Journal of Transport Geography 39, 228–237. https://doi.org/10.1016/j.jtrangeo.2014.07.013

Casas, I., Delmelle, E.C., 2017. Tweeting about public transit — Gleaning public perceptions from a social media microblog, Case Studies on Transport Policy 5 (4), 634-642. ISSN 2213-624X, https://doi.org/10.1016/j.cstp.2017.08.004.

Caunhye, A.M., Nie, X., Pokharel, S., 2012. Optimization models in emergency logistics: A literature review. Socio-Economic Planning Sciences 46 (1), 4-13. https://doi.org/10.1016/j.seps.2011.04.004

Cihon, P., Yasseri, T., 2016. A biased review of biases in twitter studies on political collective action. At the Crossroads: lessons and Challenges in Computational Social Science 91. Frontiers in physics 4. https://doi.org/10.3389/fphy.2016.00034

Chaniotakis, E., Antoniou, C., Pereira, F., 2016. Mapping Social Media for Transportation Studies. IEEE Intelligent Systems 31 (6), 64–70. https://doi.org/10.1109/MIS.2016.98

Collins, C., Hasan, S., Ukkusuri, S.V., 2013. A Novel Transit Rider Satisfaction Metric : Rider Sentiments Measured from Online Social Media Data. Journal of Public Transportation 16 (2), 21–45. https://doi.org/10.5038/2375-0901.16.2.2

Combs, T.S., Shay, E., Salvesen, D., Kolosna, C., Madeley, M., 2016. Understanding the multiple dimensions of transportation disadvantage: the case of rural North Carolina. Case Studies on Transport Policy 4 (2), 68–77. https://doi.org/10.1016/j.cstp.2016.02.004

Costantino,G., La Marra, A., Martinelli, F., Saracino, A., Sheikhalishahi, M., 2017. Privacy-preserving text mining as a service. 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 890-897. doi: 10.1109/ISCC.2017.8024639

Coughlan, P. and Coghlan, D. 2002. Action research for operations management. International Journal of Operations & Production Management 22 (2), 220-240. https://doi.org/10.1108/01443570210417515

Culotta, A., Ravi, N.K., Cutler, J., 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 72–78.

D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems 16 (4), 2269–2283. https://doi.org/10.1109/TITS.2015.2404431

Danish Regions, 2017. The future of transport – Disruption requires new flexible planning solutions; Danske Regioner, Fremtidens Transport -Disruption kræver ny fleksibel planlægning.

Elo, S., Kyngäs, H., 2008. The qualitative content analysis process. Journal of Advanced Nursing 62 (1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Elvy, J., 2014. Public participation in transport planning amongst the socially excluded: An analysis of 3rd generation local transport plans. Case Studies on Transport Policy 2 (2), 41–49. https://doi.org/10.1016/j.cstp.2014.06.004

Fadili, H., Jouis, C., 2016. Towards an Automatic Analyze and Standardization of Unstructured Data in the Context of Big and Linked Data. MEDES Proceedings of the 8th International Conference on Management of Digital EcoSystems, 223–230. https://doi.org/10.1145/3012071.3012103

Fagnant D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transportation Research Part A: Policy and Practice 77, 167-181. https://doi.org/10.1016/j.tra.2015.04.003.

Fruchterman, T.M.J., Reingold, E.M., 1991. Graph Drawing By Force-Directed Placement. Software-Practice & Experience 21 (11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Gal-Tzur, A., Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Shoor, I., 2014. The potential of social media in delivering transport policy goals. Transport Policy 32, 115–123. https://doi.org/10.1016/j.tranpol.2014.01.007

Gao, L., Wu, H., 2013. Verb-Based Text Mining of Road Crash Report. Transportation Research Board 92nd Annual Meeting, 5–16. Retrieved from http://trid.trb.org/view/2013/C/1241434

Gao, L., Yu, Y., 2016. Public Transit Customer Satisfaction Dimensions Discovery from Online Reviews. Urban Rail Transit 2, 146–152. https://doi.org/10.1007/s40864-016-0042-0

Gerlitz, C., Rieder, B., 2013. Mining One Percent of Twitter: Collections, Baselines, Sampling. M/C Journal, [S.l.], v. 16, n. 2, mar. 2013. ISSN 14412616. Available at: <http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/620>.

Grant-Muller, S.M., Kuflik, T., Minkov, E., Nocera, S., Gal-Tzur, A., Shoor, I., 2015. Transport Policy: Social Media and User-Generated Content in a Changing Information Paradigm, in Nepal, S., Paris, C., & Georgakopoulos, D. (Eds.), Social Media for Government Services, 325–365. Springer.

Gu, Y., Qian, Z.S., Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using social media data. Transportation Research Part C: Emerging Technologies 67, 321–342. https://doi.org/10.1016/j.trc.2016.02.011

Guo, Y., Barnes, S.J., Jia, Q., 2017. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management 59, 467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Jalali, S., Wohlin, C., 2012. Systematic literature studies. Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement - ESEM '12, Lund, Sweden, ACM Press. https://doi.org/10.1145/2372251.2372257

Jena, R., 2020. An empirical case study on Indian consumers' sentiment towards electric vehicles: a big data analytics approach. Industrial Marketing Management, In Press. https://doi.org/10.1016/j.indmarman.2019.12.012

J. Hall, D., R. Huscroft, J., T. Hazen, B., B. Hanna, J., 2013. Reverse logistics goals, metrics, and challenges: perspectives from industry. International Journal of Physical Distribution & Logistics Management 43 (9), 768–785. https://doi.org/10.1108/IJPDLM-02-2012-0052

Hassan, D., 2018. A Text Mining Approach for Evaluating Event Credibility on Twitter. 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, 2018, 171-174. doi: 10.1109/WETICE.2018.00039

Kim, A.E., Hansen, H.M., Murphy, J., Richards, A.K., Duke, J., Allen, J.A., 2013. Methodological Considerations in Analyzing Twitter Data. JNCI Monographs, Volume 2013 (47), 140–146. https://doi.org/10.1093/jncimonographs/lgt026

Kinra A., 2015. Environmental complexity related information for the assessment of country logistics environments: Implications for spatial transaction costs and foreign location attractiveness. Journal of Transport Geography 43, 36-47. https://doi.org/10.1016/j.jtrangeo.2014.12.005.

Kinra, A., Beheshti-Kashi, S., Pereira, F., Combes, F., Rothengatter, W., 2019. Textual data in transportation research: techniques and opportunities, in Antoniou, C., Dimitriou, L. and Pereira, F. (eds.), Mobility Patterns, Big Data and Transport Analytics, 173-197, Elsevier.

Kinra, A., Mukkamala, R.R., Vatrapu, Ravi, 2016. Methodological Demonstration of a Text Analytics Approach to Country Logistics System Assessments. Dynamics in Logistics, 119–129.

König, M., Neumayr, L., 2017. Users resistance towards radical innovations: The case of the self-driving car. Transportation Research Part F: Traffic Psychology and Behaviour 44, 42–52. https://doi.org/10.1016/j.trf.2016.10.013

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihók, G., Den Hartog, D.N., 2018. Text Mining in Organizational Research. Organizational Research Methods 21 (3), 733–765. https://doi.org/10.1177/1094428117722619

Kosala, R., Adi, E., Steven, (2012). Harvesting real time traffic information from twitter. Procedia Engineering 50, 1–11. https://doi.org/10.1016/j.proeng.2012.10.001

Krippendorff, K., 2013. Content analysis : An Introduction to its Methodology (3rd ed.). SAGE Publications.

Kyriakidis, M., Happee, R., de Winter, J.C.F., 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. Transportation Research Part F: Traffic Psychology and Behaviour 32, 127-140. doi: 10.1016/j.trf.2015.04.014.

Kumar, S., Morstatter, F., Liu, H., 2014. Twitter Data Analytics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-9372-3

Kumar, A., Monica, M., Rinkita, M., 2016. Big Data on Content Credibility of Social Networking Sites and Instant Messaging Applications. IOSR Journal of Computer Engineering 18, 27-31. doi: 10.9790/0661-1805052731

Kühl, N., Goutier, M., Ensslen, A., Jochem, P. 2019. Literature vs. Twitter: empirical insights on customer needs in e-mobility. J. Cleaner Prod., 213 , 508-520. https://doi: 10.1016/j.jclepro.2018.12.003

Lee, R., Sener, I., 2015. Transportation and Quality of Life: Where Do They Intersect? Journal of Transport & Health 2 (2), 77–78. https://doi.org/10.1016/j.jth.2015.04.487

Lomborg, S., 2016. A state of flux: histories of social media research. European Journal of Communication 32 (1), 6–15. https://doi.org/10.1177/0267323116682807

Lomborg, S., Bechmann, A., 2014. Using APIs for Data Collection on Social Media. The Information Society 30 (4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Luong, T.T.B., Houston, D., 2015. Public opinions of light rail service in Los Angeles, an analysis using Twitter data. iConference 2015 Proceedings, 2–5.

Maghrebi, M., Abbasi, A., Rashidi, T., Waller, S. 2015. Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 208-213. doi: 10.1109/ITSC.2015.43.

Medhat, W., Hassan, A., Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal 5 (4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., Nisbet, R.A., 2012. Practical Text Mining and Statistical Analysis for Non -structured Text Data Applications. Academic Press.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., Rosenquist, J.N., 2011. Understanding the Demographics of Twitter Users. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 554–557. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Morstatter, F., Kumar, S., Liu, H., Maciejewski, R., 2013a. Understanding Twitter data with TweetXplorer. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13. New York, NY: ACM. doi: 10.1145/2487575.2487703

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., 2013b. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. New York, NY: AAAI Press.

Näslund, D. 2002. Logistics needs qualitative research–especially action research. International Journal of Physical Distribution & Logistics Management 32 (5), 321-338. https://doi.org/10.1108/09600030210434143

Näslund, D., Kale, R. and Paulraj, A. 2010. Action research in supply chain management–a framework for relevant and rigorous research. Journal of Business Logistics 31 (2), 331-355.

Newnam, S., Goode, N., Salmon, P., Stevenson, M., 2017. Reforming the road freight transportation system using systems thinking: An investigation of Coronial inquests in Australia. Accident Analysis & Prevention 101, 28–36. https://doi.org/10.1016/j.aap.2017.01.016

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2 (1–2), 1–135. doi: 10.1561/1500000011

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2013. Text analysis in incident duration prediction. Transportation Research Part C: Emerging Technologies 37, 177–192. https://doi.org/10.1016/j.trc.2013.10.002

Pfeffer, J., Mayer, K., Morstatter, F., 2018. Tampering with Twitter's Sample API. EPJ Data Science 7, 50. https://doi.org/10.1140/epjds/s13688-018-0178-0

Pokharel, S., Mutha, A., 2009. Perspectives in reverse logistics: A review. Resources, Conservation and Recycling, 53 (4), 175–182. https://doi.org/10.1016/j.resconrec.2008.11.006

Rabinovich, E., Cheon, S., 2011. Expanding horizons and deepening understanding via the use of secondary data sources. Journal of Business Logistics 32 (4), 303–316. https://doi.org/10.1111/j.0000-0000.2011.01026.x

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. Transportation Research Part C: Emerging Technologies 75, 197–211. https://doi.org/10.1016/j.trc.2016.12.008

Riffe, D., Lacy, S., Fico, F. 2005. Analyzing Media Messages: Using Quantitative Analysis in Research. Mahwah, NJ: Lawrence Erlbaum Associates.

Rodrigues, F., Borysov, S.S., Ribeiro, B., Pereira, F.C. 2017. A Bayesian additive model for understanding public transport usage in special events. IEEE transactions on pattern analysis and machine intelligence 39 (11), 2113-2126.

Ryley, T., Gjersoe, N., 2006. Newspaper response to the Edinburgh congestion charging proposals. Transport Policy, 13 (1), 66–73. https://doi.org/10.1016/j.tranpol.2005.08.004

Schoettle, B., Sivak, M., 2014. A survey of public opinion about autonomous and self-driving vehicles in the U.S., The U.K., and Australia. The University of Michigan, Transportation Research Institute, Michigan.

Schulz, A., Ristoski, P., Paulheim, H., 2013. I see a car crash: Real-time detection of small scale incidents in microblogs. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7955, 22–33. https://doi.org/10.1007/978-3-642-41242-4_3

Schweitzer, L., 2014. Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter. Journal of the American Planning Association 80 (3), 218–238. https://doi.org/10.1080/01944363.2014.980439

Spens, K.M., Kovács, G., 2006. A content analysis of research approaches in logistics research. International Journal of Physical Distribution & Logistics Management 36 (5), 374–390. https://doi.org/10.1108/09600030610676259

Stone, P.J., Dunphy, D.C., Smith, M.S., 1966. The general inquirer: A computer approach to content analysis. Oxford, England: M.I.T. Press.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., 2010. Sentiment Strength Detection in Short Informal Text. The American Society for Informational Science and Technology 61 (12), 2544–2558. https://doi.org/10.1002/asi

Tirunillai, S., Tellis, G.J., 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research 51 (4), 463–479. https://doi.org/10.1509/jmr.12.0106

Tufekci, Z., 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Twitter (2011). One hundred million voices. Retrieved from: https://blog.twitter.com/2011/one-hundred-million-voices, April 10, 2017.

Varol, O., Ferrara, E., Davis, C., Menczer, F., Flammini, A., 2017. Online Human-Bot Interactions: Detection, Estimation, and Characterization. International AAAI Conference on Web and Social Media, North America. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817. Date accessed: 30 Sep. 2019.

Vejdirektoratet, Wilke, 2017. Danish Expectations For Self-Driving cars OR Danskernes forventninger til selvkørende biler. http://www.vejdirektoratet.dk/DA/viden_og_data/temaer/Selvkoerendebiler/Documents/Rapport_070217_short.pdf

Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 186-194.

Wang, F., Xu, T.H., Zhao, Y., Huang, Y.R., 2015. Prior LDA and SVM Based Fault Diagnosis of Vehicle On-board Equipment for High Speed Railway. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, 818–823. https://doi.org/10.1109/ITSC.2015.138

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., Chaovalit, P., 2011. Social-based traffic information extraction and classification. 2011 11th International Conference on ITS Telecommunications, 107–112. https://doi.org/10.1109/ITST.2011.6060036

Webb, L., Wang, Y., 2013. Techniques for sampling online text-based data sets, in Hu, W.C., Kaabouch, N. (eds.), Big data management, technologies, and applications, IGI Global Publishers. 10.4018/978-1-4666-4699-5.ch005.

Williams, T., Betak, J., Findley, B., 2016. Text Mining Analysis of Railroad Accident Investigation Reports. 2016 Joint Rail Conference. American Society of Mechanical Engineers. https://doi.org/10.1115/JRC2016-5757

Xuan, L., El-Gohary, N., 2016. Text Analytics for Supporting Stakeholder Opinion Mining for Large-scale Highway Projects. Procedia Engineering 145, 518–524. https://doi.org/10.1016/j.proeng.2016.04.039

Zanini, N., Dhawan, V., 2015. Text Mining: An introduction to theory and some applications. Research Matters 19, 38-44.