Understanding the Public Opinion of Automated Vehicles in Pittsburg Using Structural Topic Modeling

Varshini Kamaraj

University of Wisconsin - Madison

Abstract

In September 2016 Pittsburg became ground zero for the testing of Uber's fleet of automated vehicles(AVs). Pittsburg is also home to BikePGH, a local non-profit that advocates for walking and biking friendly streets. To understand the public opinion on the testing of AVs in Pittsburg, BikePGH conducted a survey to understand how pedestrians and bicyclists felt about sharing the streets with AVs. This survey consisted of both open-ended responses and close-ended responses and was made publicly available.

Survey data consisting of close-ended responses can be easily analyzed. However, open-ended comments are tedious to analyze as they are typically analyzed manually. These open-ended responses can give researchers useful insights into public opinion. Such insights could possibly shape how AVs are designed, tested and how policies regarding the deployment of AVs are drafted. In order to overcome the manual analysis of the open-ended survey data, machine learning methods for content analysis of text data can be used. Structural topic modeling(STM) is one such popular method that allows researchers to discover topics in text data. Structural topic modeling can be conducted using STM, an R package that has functions for conducting each stage of structural topic modeling.

An analysis of the open-ended survey responses using the STM package in R is presented here. Topic modeling was performed on responses to the question: "What do you think BikePGH's position on AVs should be? Please elaborate on this." Various other R packages such as ggplot2 along with the STM package are used for data analysis and visualization. The STM package extracted six topics from the survey data. From the results of the topic model, it could be concluded that the public was either supportive or non-supportive of AVs. Some opined that BikePGH should take a position regarding the use of AVs while others stated that BikePGH should adhere to creating a biking/walking friendly neighborhoods. Finally, some topics seem to indicate that a part of the public feels that AVs are safer than regular cars and that AVs have the potential to reduce risk. The STM model for the BikePGH data set can be found in a GitHub repository: https://github.com/avkamaraj/Final_Integration_PittsburgData

Introduction

BikePGH is a non-profit group based in Pittsburg that advocates for biking and walking friendly streets in Pittsburg. In September 2016 Uber made Pittsburg the first place where ride-share users could use a self-driving vehicle. Upon commencement of the automated vehicle(AV) ride-sharing service BikePGH noticed an influx of comments/complaints from people who were interacting with AVs on the streets of Pittsburg. To better understand the public opinion on AVs and to initiate a conversation about the use of AVs and its testing, BikePGH conducted a survey to understand how pedestrians and bicyclists felt about sharing the streets with AVs.

The survey consisted of both open-ended and closed-ended questions. Open-ended questions give responders the opportunity to answer in sentences and express their thoughts. These responses often give more insight into the problem being researched. Closed-ended questions can be answered by selecting an answer from a list of possible answers. For example, answers like "Yes" or "No" or ratings from 1-5.

An example of an open-ended question in the survey is: “What do you think BikePGH's position on AVs should be? Please elaborate on this”. Analyzing such questions can give developers of AVs and policymakers a better understanding of the public opinion of AVs and potentially give rise to changes in the design and operation of AVs.

Open-ended responses can be used to gain a better understanding of a responder’s thoughts on a specific subject matter. Manual analysis of open-ended responses and discovering themes among the responses is a tedious process. Machine learning methods can be used for analyzing such data and for identifying patterns within the text. Since the goal is to identify 'patterns' in any given a dataset, unsupervised learning methods can be used for analysis.

Unsupervised learning is a type of machine learning algorithm that can be used to draw inferences from data sets consisting of input data without labeled responses. Since the goal is to identify themes/topics from the text data, topic modeling is proposed for analyzing survey data. Topic models are used to make inferences about 'topics' in a text document. It is an unsupervised learning procedure that works similar to clustering.

The topic modeling method proposed here for analyzing the BikePGH dataset is structural topic modeling(STM). Structural topic models are said to be an improvement over text analysis methods such as Latent Dirichlet Allocation(LDA) and Correlated Topic Model(CTM). This is because STM allows users to include documents metadata into the topic model. For analysis of the BikePGH data set, the structural topic modeling is conducted using *STM*, an R package that has functions for conducting each stage of structural topic modeling. Several other R packages like *tidyverse* are used along with the STM package to aid in the data analysis and visualization.

Methods

The BikePGH survey was conducted by in two parts. One survey was conducted exclusively for donor-members and received 321 responses. Another survey was conducted to receive responses from the public and this received 798 responses. The survey consisted of both open-ended and closed-ended questions. The goal of using structural topic modeling here is to understand the public opinion on the introduction of automated vehicles into their city. Thus, the modeling is focused on using the survey data from the public.

The first analysis is that of the close-ended responses. There was a total of 23 variables in the data set. A data dictionary explaining the variables can be found on the BikePGH website. In order to clean the data, the names of the variables were changed to be the same as the variables names listed in the data dictionary. Variables that were unnecessary to the analysis like start date, end date, zip code were removed. For analyzing the close-ended comments, the open-ended comments were filtered out as well. Finally, fields, where the responses were blank, were also removed. After cleaning the dataset, there were 14 variables with 793 observations. Count and mutate operations from the *dplyr* R package were used to find the percentage of the responses to a question.

The next analysis involves the open-ended responses. These responses also need to be in a tidy format before conducting the topic modeling. The STM package includes functions that can be used for cleaning the responses. First, the excel file containing the responses are imported into R. The STM package consists of a function called textProcessor() which can be used to process the text data. The pre-processing of the text data typically includes stemming (reducing works to their root form), dropping punctuations and stop word removal (removing words like 'the.' The textProcessor() function is capable of performing all three steps. Another step for cleaning the open-ended responses involves removing words that do not appear frequently in the text. This can be achieved using the prepDocuments() function by setting a parameter named lower.thresh. Before setting the lower threshold value another function called plotRemoved() can be used to plot the number of words and documents that will be removed for different threshold values.

Note that for conducting the analysis of the open-ended responses, the focus was only on one question. There was a total of four open-ended questions in the survey. Since the goal was to understand public opinion, the closest question that could represent public opinion was determined to be “What do you think BikePGH's position on AVs should be? Please elaborate on this.”

Results

Close-Ended Survey

Figure 1, Figure 2, Figure 3 and Figure 4 depict the public response to the close-ended survey response. It can be observed that about 56% and 47% of responders have interacted with automated vehicles either while walking or while riding bikes. It can also be seen that only 5% of the responders felt very safe using Pittsburg streets with human-driven cars, while 19% report that they feel very safe using Pittsburg streets with autonomous vehicles.

The public was also surveyed regarding several regulatory issues. These questions asked "On public streets, do you think that a regulatory authority should…(a) Come up with regulations regarding how AVs are tested? (b) Cap the speed limit in which AVs are allowed to operate? (c) Prevent AVs from operating in an active school zone? (d) Require companies to share non-personal data with the proper authorities, planning agencies, etc?

70 % of responders answered 'Yes' for regulations on how AVs are tested. 53% voted 'Yes' for capping the speed limit in which AVs are allowed to operate. 46% were against the use of AVs in an active school zone and 71% agree that regulations should require companies to share non-personal data with proper authorities. In terms of welcoming AVs to Pittsburg for testing, 49% of the responders approve the usage of Pittsburg as a proving ground for AVs and 43% feel that BikePGH should actively support AVs. The survey also enquired if BikePGh should dedicate advocacy resources to AVs for which 33% agreed. The public was also asked about their familiarity with AVs on the news and their familiarity with the technology that drives automated vehicles. 33% responded that they paid attention to the news concerning AVs to 'a moderate extent' while 42% responded that they were 'somewhat familiar' with the technology behind AVs.

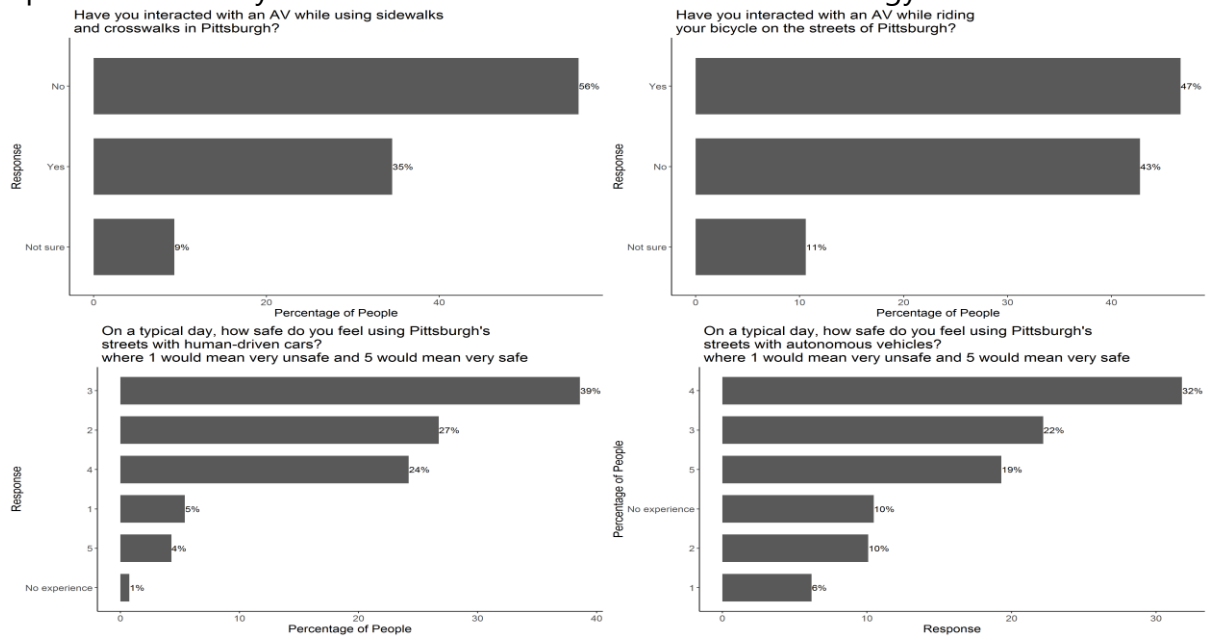

Figure 1. Public experience with automated vehicles

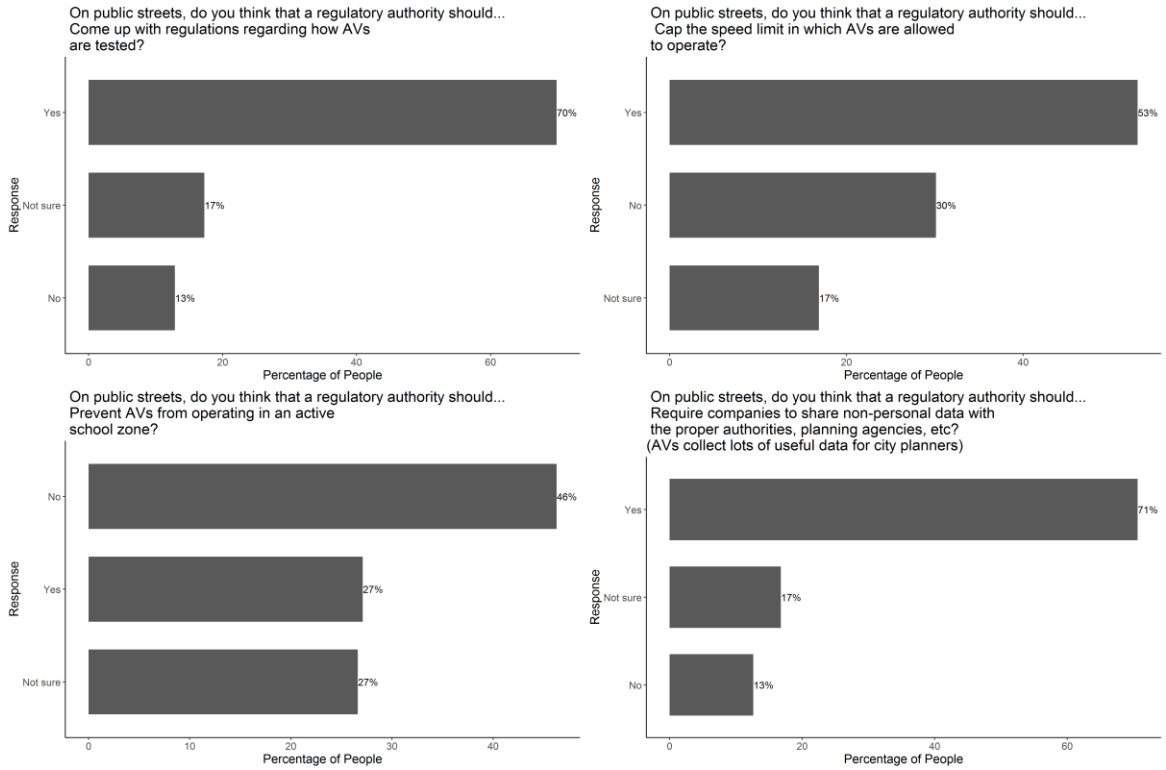

Figure 2. Public views on regulations regarding automated vehicles

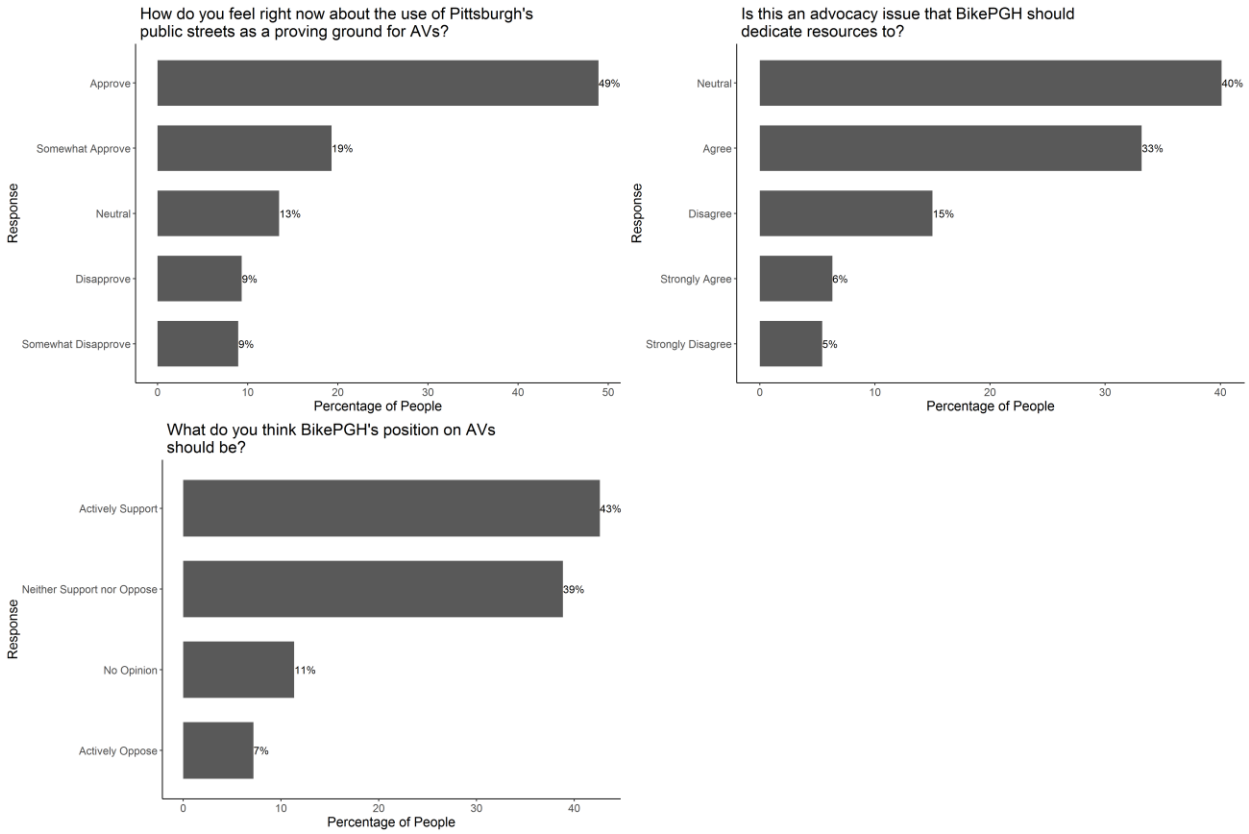

Figure 3. Views on advocacy

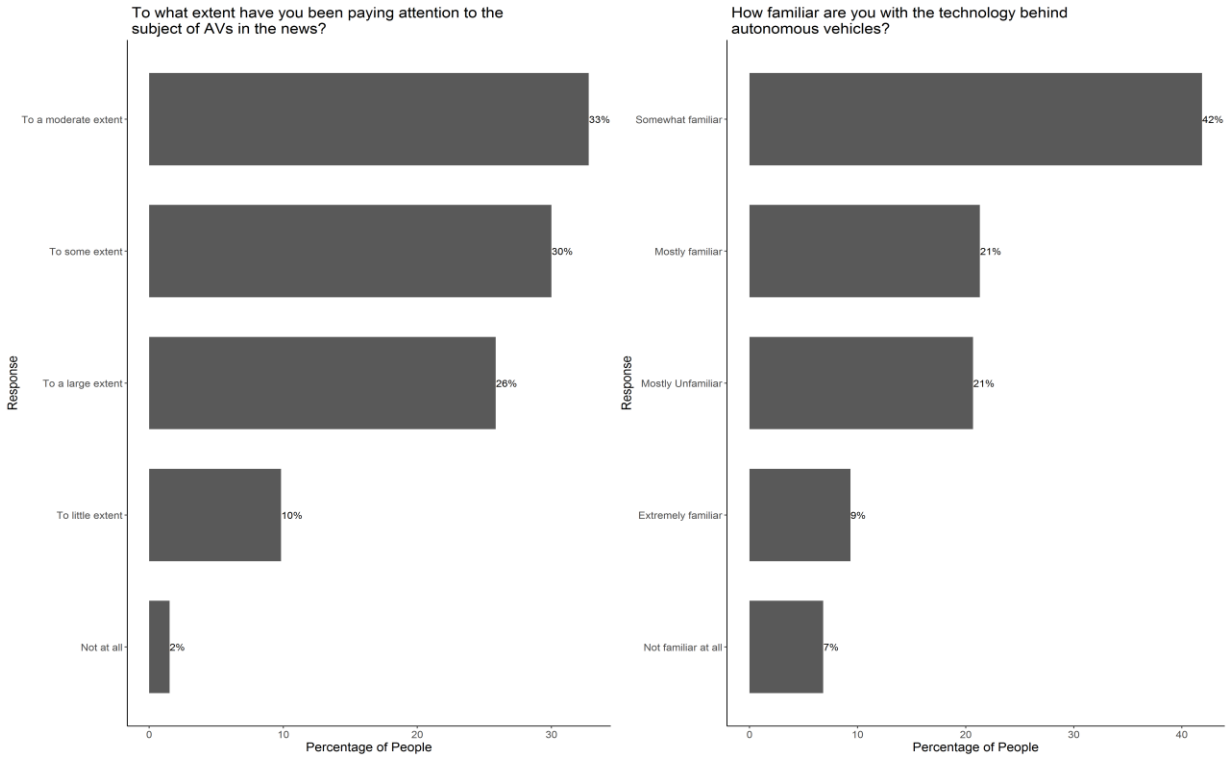

Figure 4. Familiarity with automated vehicles

Open-Ended Survey: Topic Selection and Topic Content

Structural topic modeling is an unsupervised learning method, and this means that the number of topics is user-defined. There exist criteria that can guide user selection of the number of topics. The selection criteria include held-out likelihood, residual variance, coherence, and exclusivity. It is important to remember that while these criteria will give an optimum number for the number of topics, it is up to the analyst to ultimately decide if these topics make sense. The STM package offers a searchK() function that can be used to run through different user-specified topic numbers and evaluate the four selection criteria. The models considered ranged from K=5 to K=12 where K is the number of topics. Figure 5 shows the Pareto frontiers for these models. The points that are below the lines represent inferior models. Based on the Pareto frontiers, K =5 and K=6 were considered for the number of topics. Upon running the analysis, the interpretability of the topics appeared to be better with 6-topic models. Thus, a 6-topic model was selected.

Once the number of topics was selected, the meaning of the topics was identified by looking at the corresponding comments from which the terms were extracted. Table 1 shows the terms that identify the topic and the corresponding comments associated with the topic. In Table 1 the term *Prob* refers to the probability that a term occurs in a topic. *Frex* identifies terms that occur frequently in a topic and are exclusive to that topic. *Lift* weights words more heavily if they occur infrequently in other topics and *score* weights words by dividing the logarithm of their frequency by the logarithm of their frequency in other topics.

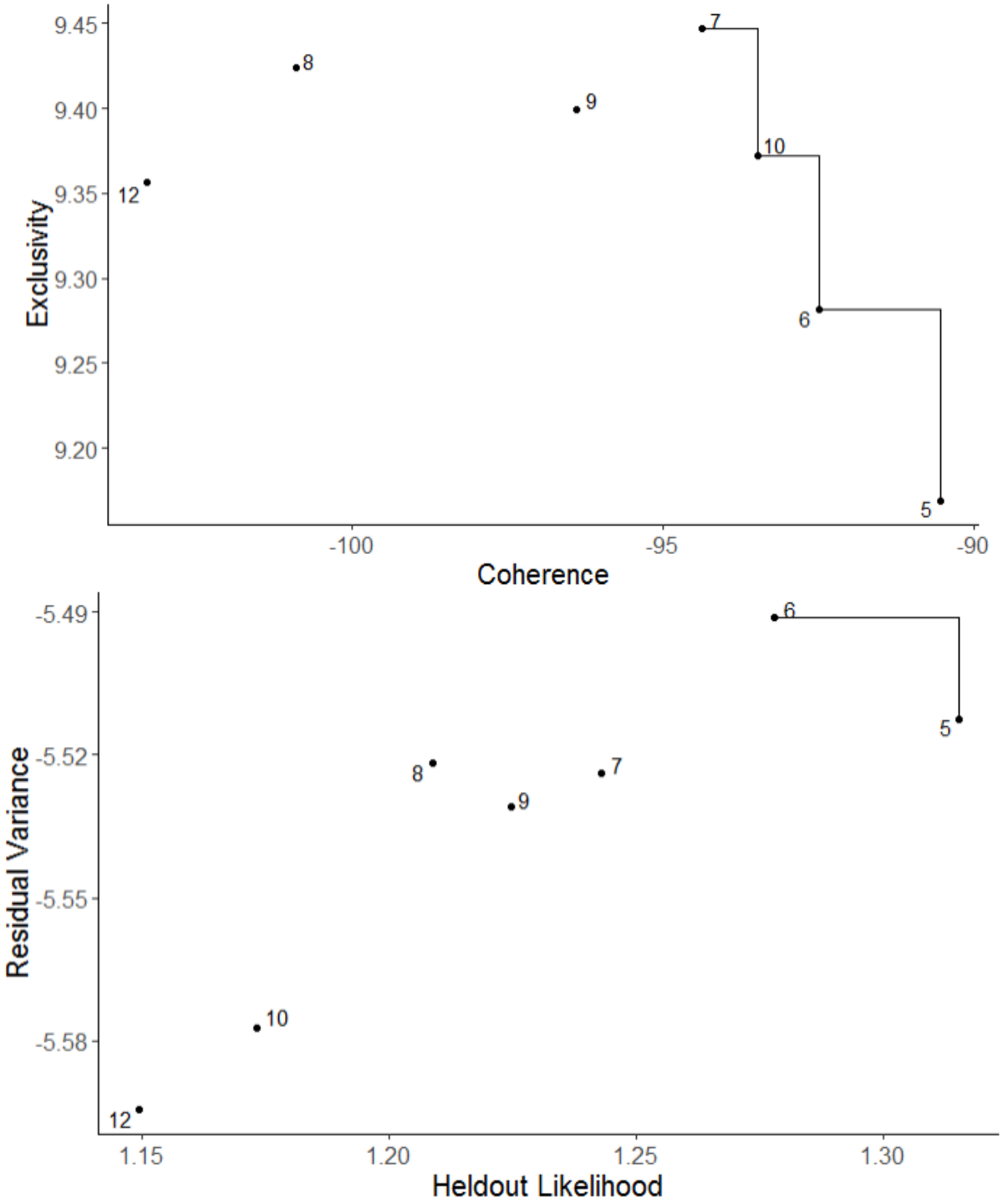

Figure 5. Pareto frontiers for topic selection

Table 1. Topic and Relevant Comments

Topic 1: Non-supportive/Too soon to support • Highest Probability: vehicl, street, dont, autonom, take, seem, oper • FREX: vehicl, hand, import, seem, dont, problem, oper • Lift: collect, hand, motor, replac, tell, track, wait • Score: vehicl, tell, dont, autonom, oper, street, seem	The jury and data are still out on how these vehicles will operate and if they can safely operate with pedestrians and cyclists. I would say that bike Pittsburgh should support the ongoing research and data collection. But currently coming out in favor or against these vehicles seems too soon. That is based on my opinion with limited knowledge of the current research that exists.
Topic 2: Work with Technology • Highest Probability: can, public, need, work, safe, oppos, advoc • FREX: public, oppos, advoc, data, posit, experi, live • Lift: anyon, easier, effort, engag, experi, fight, invest • Score: next, data, advoc, oppos, public, work, interest	I feel that this is the future and we need to work to get along, even with non-humans. If we work together in the start then more positive result follow. To go head in anti anything new or advanced than what do we have to gain? As we age we are less open minded and open to new things. Lets reverse that and embrace the future and work side by side. Maybe they can put a bike rack on the back someday.
Topic 3: Neutral • Highest Probability: safeti, test, issu, develop, pedestrian, regul, potenti • FREX: issu, develop, test, safeti, turn, reduc, continu • Lift: allow, ask, avoid, cours, develop, failur, input • Score: turn, test, safeti, develop, issu, regul, reduc	If done correctly, AVs have the potential to reduce pedestrian and cyclist injuries/fatalities. However, during the testing phase they may increase. Therefore I would take a neutral stance.
Topic 4: AVs are safe • Highest Probability: driver, car, human, safer, drive, pedestrian, will • FREX: driver, human, drive, distract, around, far, hit • Lift: alway, close, cross, decreas, distract, far, ive • Score: driver, human, drive, monitor, distract, phone, hit	AVs are much safer than human drivers. The more AVs are on the road, the safer they are for bikers, pedestrians and other drivers.
Topic 5: AVs safe if done correctly • Highest Probability: think, will, make, bikepgh, cyclist, support, safer • FREX: sure, involv, think, everyon, happen, make, might • Lift: everyon, determin, differ, goal, happen, might, opportune • Score: think, bikepgh, will, protect, cyclist, make, sure	AVs have the opportunity to increase safety for cyclists, if done correctly. I think BikePGH should stay actively involved to be sure cyclists are being considered in practical and helpful ways- also to accurately communicate what is happening to the cycling community.
Topic 6: Anti-AV • Highest Probability: bike, citi, pittsburgh, pgh, bicyclist, transport, need • FREX: bike, just, lane, focus, group, citi, transport • Lift: infrastructur, money, polit, realiti, bike, area, first • Score: bike, practic, pgh, transport, citi, part, infrastructur	bike pgh should focus more on the proposed bike lanes and adding more flexibility to ride throughout the city

Discussion

Topics Identified

The algorithm identified 6 topics. Upon assessing the words and the associated comments it can be observed (according to the algorithm) that the public takes one of the following stances,

- They are either non-supportive of AVs or feel it is too soon to support AVs.
- They believe the future is with autonomous vehicles and that technology should be embraced
- They take a neutral stance possibly due to fatalities during testing
- They believe AVs are far safer than humans
- They support AVs and encourage BikePGH to advocate for and communicate progress with the biking community
- They do not support AVs and want BikePGH to stick to advocating for biking/walking.

Algorithm reliability, validity, and limitations

In evaluating the algorithm, it is important to assess if the algorithm is reliable and valid. Here reliability implies if the model is repeatable. In order to get repeatable results, a seed can be set in the algorithm to get repeatable results. Furthermore, it would be beneficial to set different seeds and see if the model arrives at similar kind of topics. Validity refers to the correctness of the model. This brings us to one of the limitations of topic modeling. Topic modeling cannot be used to arrive at an accurate number of topics to describe the data. For instance, the number of topics is left to the user to decide. In the STM package, the function searchK() can be used to select the number of topics. However, these functions are only data-driven approaches that are intended to guide the user and not give a definitive answer for the number of topics. Thus, it is necessary for the user to have a theoretical understanding of the data while deciding on the number of topics.

Another drawback in this model is the way in which the raw data is preprocessed. Note that the methods and order in which the preprocessing of the raw text is performed could possibly change the results. For example, here the text is preprocessed using a function available in the stm package called the textProcessor(), the working of this function isn't readily clear to the user. This could be a source of variability in the results.

Conclusions

Structural topic modeling is a powerful and convenient method for analyzing substantial amounts of text data. It can be used for testing numerous hypotheses regarding public perception on different issues based on the results obtained from topic modeling. Although the method has a variety of advantages in terms of analysis, it is important to bear in mind that topic modeling has the potential to generate results that can drive policymaking in different areas. For example, setting a smaller value for the number of topics extracted could result in a binary outcome that is probably not a reflection of the

nature of the data being analyzed. Such errors could drive environmental policies or maybe be suggestive of opinions that either advance or hinder the development of several types of technologies. Thus, even though STM is a powerful tool for content analysis, it is imperative that algorithm transparency, reliability and validity are ensured as much as possible.

References

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013, January). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation* (pp. 1-20).

Data Catalog. (n.d.). Retrieved December 21, 2018, from https://catalog.data.gov/dataset?tags=bikepgh

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".