


Fall 2018

## Evaluation of public feedback for the Interstate-80 Planning Study in Iowa.

Sai Saketh Katangur  
Iowa State University

Follow this and additional works at: <https://lib.dr.iastate.edu/creativecomponents>

 Part of the [Business Analytics Commons](#), [Civic and Community Engagement Commons](#), [Community-Based Learning Commons](#), and the [Community-Based Research Commons](#)

### Recommended Citation

Katangur, Sai Saketh, "Evaluation of public feedback for the Interstate-80 Planning Study in Iowa." (2018).  
*Creative Components*. 70.  
<https://lib.dr.iastate.edu/creativecomponents/70>

This Creative Component is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Creative Components by an authorized administrator of Iowa State University Digital Repository. For more information, please contact [digirep@iastate.edu](mailto:digirep@iastate.edu).

**Evaluation of public feedback for the Interstate-80 Planning Study in Iowa.**

by

**Sai Saketh Katangur**

A Creative Component report submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

Master of Science.

Major: Information Systems

Program of Study Committee:

Dr. Nilakanta Sree

Dr. James Davis A

The student author, whose presentation was approved by the program of study committee, is solely responsible for the content of this report. The Graduate College will ensure this report is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

Copyright © Sai Saketh Katangur, 2018. All rights reserved.

## Table of Contents

LIST OF FIGURES .....	3
ABBREVIATIONS .....	5
ACKNOWLEDGMENTS .....	6
ABSTRACT .....	7
INTRODUCTION .....	8
LITERATURE REVIEW .....	10
Text Mining.....	10
Topic Modeling .....	10
Latent Dirichlet Allocation.....	10
METHODOLOGY .....	12
Overview .....	12
Survey Design and Implementation .....	12
RESULTS .....	14
Evaluation and Analysis of Quantitate Data .....	14
Evaluation and Analysis of Qualitative Data .....	23
LIMITATIONS.....	35
CONCLUSION.....	36
REFERENCES .....	37
APPENDIX A – R Code.....	38
APPENDIX B – Survey Questionnaire .....	50

## LIST OF FIGURES

	Page
<i>Figure 1: Gender Distribution</i> .....	14
<i>Figure 2: Age Distribution</i> .....	14
<i>Figure 3: Average weekly miles driven on I-80</i> .....	15
<i>Figure 4: Frequency of Travel on I-80</i> .....	15
<i>Figure 5: Purpose of travel on I-80</i> .....	16
<i>Figure 6: Level of traffic on I-80</i> .....	17
<i>Figure 7: Current I-80 conditions - 1</i> .....	17
<i>Figure 8: Current I-80 conditions - 2</i> .....	18
<i>Figure 9: Importance of future I-80 conditions</i> .....	18
<i>Figure 10: Traffic Congestion on I-80 in next 10 years</i> .....	19
<i>Figure 11: Condition that would result in taking alternative routes</i> .....	20
<i>Figure 12: Importance of adding lanes to I-80</i> .....	20
<i>Figure 13: People's opinions on various factors</i> .....	21
<i>Figure 14: Level of Urgency</i> .....	21
<i>Figure 15: Importance of having three lanes on the rural interstate system</i> .....	22
<i>Figure 16: Overall Interstate System rating</i> .....	22
<i>Figure 17: Preliminary LDA Results - Positive Impact</i> .....	25
<i>Figure 18: LDA results -without custom words - Positive Impact</i> .....	26
<i>Figure 19: LDA Results - Bigrams - Positive Impact</i> .....	27
<i>Figure 20: LDA results - Trigrams - Positive Impact</i> .....	28
<i>Figure 21: Bigram Frequencies - Positive Impact</i> .....	29
<i>Figure 22: Trigram frequencies - Positive Impact</i> .....	29

<i>Figure 23: Preliminary LDA results - Negative Impact</i> .....	30
<i>Figure 24: LDA results - without custom words - Negative Impact</i> .....	31
<i>Figure 25: LDA results - Bigrams - Negative Impact</i> .....	32
<i>Figure 26: LDA Results - Trigrams - Negative Impact</i> .....	33
<i>Figure 27: Bigram Frequencies - Negative Impact</i> .....	34
<i>Figure 28: Trigram Frequencies - Negative Impact</i> .....	34

Page

**ABBREVIATIONS**

DOT	Department of Transportation
I-80	Interstate 80
PEL	Planning Environmental Linkages
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation

## ACKNOWLEDGMENTS

I have put efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am ineffably indebted to our Head of the Committee and Major Professor Dr. Sree Nilakanta, for his guidance and constant supervision as well as providing me with ideas that gave me direction for working on and completing this project.

I am extremely thankful and pay my gratitude to my faculty and Academic Advisor, Dr. James Davis for his valuable guidance and support throughout my masters program.

My sincere thanks to my manager, Mr. Matthew Haubrich for providing me with the opportunity, resources and guidance throughout this project.

I also acknowledge with a deep sense of reverence, my gratitude towards my parents and members of my family, who has always supported me morally as well as economically.

At last but not least gratitude goes to all of my friends who directly or indirectly helped me to complete this project report.

Any omission in this brief acknowledgement does not mean lack of gratitude.

## ABSTRACT

This report evaluates the results of a survey that was conducted by the Iowa Department of Transportation to study the existing conditions, the way the system is performing, short and long-term issues and strategies to improve the I-80 in Iowa from Illinois to Nebraska borders. The goal of this component of the I-80 study is to gather feedback from the general public and evaluate, visualize the results which would be vital for the management while making decisions about the investment and development plans of the I-80.

After the analysis of the quantitative component of the survey, we could clearly display the public opinions through a series of visualizations which gives the management the required information to consider in the decision making process. The qualitative component of the survey was analyzed using Natural Language Processing by implementing the Latent Dirichlet Allocation algorithm. Although the results of the qualitative analysis were not conclusive as we would have expected, they still give an idea of the various steps that could be taken by the DOT which would have positive and negative impacts on the public.

**Keywords:** Open-ended survey analysis, LDA, Iowa DOT, Interstate.



## INTRODUCTION

Roadways provide access and the ability to move people and goods from place to place. The Interstate system was built to maximize the mobility across states. However, the Traffic volumes on the Interstate system have been growing over the years and the growth is expected to continue. With the increasing traffic volumes, several technologies are also being developed with the potential to better alter our transportation system. In one such attempt to improve mobility across the Interstate system, the Iowa Department of Transportation is studying the Interstate 80 Corridor from Council Bluffs in western Iowa to Bettendorf in eastern Iowa to evaluate safety, capacity Infrastructure deficiencies. Nationally, I-80 extends nearly 3,000 miles from California to New Jersey. In Iowa, the I-80 spans over 300 miles carrying vehicles volumes ranging from 20,000 to 35,000 vehicles per day with huge truck/freight traffic making up anywhere from 15 to 30 percent of the traffic. Interstate 80 in Iowa is vital to the state and national economy providing the infrastructure to move people and goods across Iowa and throughout the nation. The purpose of the Planning Study is to determine whether the current infrastructure will meet the demands in the next 30 years and what potential improvement alternatives will be necessary in the foreseeable future. The study will follow the Planning Environmental Linkages model that allows the Iowa DOT to establish a vision and goals for the system and also give the opportunity to study several improvement strategies early in the planning process. The study will also help the department prioritize components of the Interstate for further development.

Since transportation projects can greatly affect a community, public input is extremely important. Public involvement allows interested individuals the chance to provide ideas and comments regarding the development of a transportation project. The I-80 corridor study's success hinges on communication and cooperation with the public, local communities, state and

federal agencies. It was determined that the most efficient way to reach out to a larger audience would be through online information. A project website was created to display information regarding the study so that one consistent message is being distributed to the general public and stakeholders and also allows them to provide their feedback and share their opinions with the DOT. This project is one component of the entire I-80 and focuses on learning the public opinions about the current infrastructure, strategies and the positive/negative impacts of various changes or improvements that could be implemented by the DOT. For this purpose, a survey was conducted to learn the public opinions on important factors that were important for the I-80 study.

## LITERATURE REVIEW

### Text Mining

Text Mining was first proposed in 1995 by Ronen Feldman et al, which was then described as “The Process of extracting interesting Patterns from very large text collections for the purpose of discovering knowledge [1].” The whole idea being text mining is to extract knowledge/information from huge mass of text without having to manually go through each line. This is inarguably a tough task for a computer to perform considering how it is difficult for them to understand tones, grammar and other abstract concepts like sarcasm which human beings employ in their speech and their writings. Text Mining as such is a very broad term used for various sub techniques like text clustering, text classification, sentiment analysis, text summarization, correlation analysis, distribution analysis, information extraction, theme extraction etc [2][3][4].

### Topic Modeling

Topic modeling is one of the most powerful techniques in text mining for information extraction, finding hidden structures/semantics/relationships within the data and text documents. There has been a lot of research and many published articles in this field and it is applied in various fields. There are various probabilistic topic modeling and unsupervised machine learning algorithms that were developed which could extract knowledge from text automatically.

### Latent Dirichlet Allocation

LDA is an unsupervised generative probabilistic method for modeling a corpus and it is a very popular algorithm in topic modeling. It was first introduced by Blei, Ng and Jordan in 2003[5]. “The basic idea is that the documents are represented as random mixtures over latent topics, where

a topic is characterized by a distribution over words.” [6]. So, LDA represents topics by word probabilities. Each document is modeled as a mixture of topics, and each topic is a discrete probability distribution that defines how likely each word is to appear in a given topic. These topic probabilities give a concise representation of a document. The words with highest probabilities in each topic usually give a good idea of what the topic could have been built around. Here, a "document" is a "bag of words" with no structure beyond the topic and word statistic. [7]. LDA has been used for various open-ended survey question analysis and should ideally work well in our case where it considers each survey response as a document and should be able to generate topics with words and their probabilities of belonging to that topic.

## METHODOLOGY

### Overview

This study used a structured survey that comprised of questions with fixed response options and questions with open ended response options. The goal is to analyze the survey responses and display the results in a way that the management at DOT could clearly visualize and factor in the public opinions in their development plans for the I-80. Since the number of survey responses is not huge, the basic evaluation and descriptive analysis could be done and visualized in MS Excel itself without the need of using advanced Analytics tools or programming languages. After a brief research about the available tools and techniques for open ended survey Analysis, theme extraction through NLP using the Latent Dirichlet Allocation algorithm seemed like a good option to extract information from the open-ended survey questions. The LDA algorithm and associated data cleaning and data transforming is going to be implemented in R, which is a programming language and free software environment for statistical computing and displaying graphics.

### Survey Design and Implementation

The study employed a questionnaire consisting of a total of 20 questions which was designed on the SurveyMonkey website. The first few questions requested information like Gender, Age, Zip Code, Hours of travel on the Interstate, Miles of travel on the Interstate, Purpose of travel on the Interstate etc. and there were questions requesting the survey respondent's opinions on the current level of traffic on the I-80, their level of satisfaction with the current conditions of the I-80, importance of adding lanes to the I-80 and other fixed response questions each of whose results will be displayed in the later sections of the report.

The survey also had 2 open ended questions:

1. What changes or improvements to interstate transportation would impact you/your life in a positive way?

2. What changes or improvements to interstate transportation would impact you/your life in a negative way?

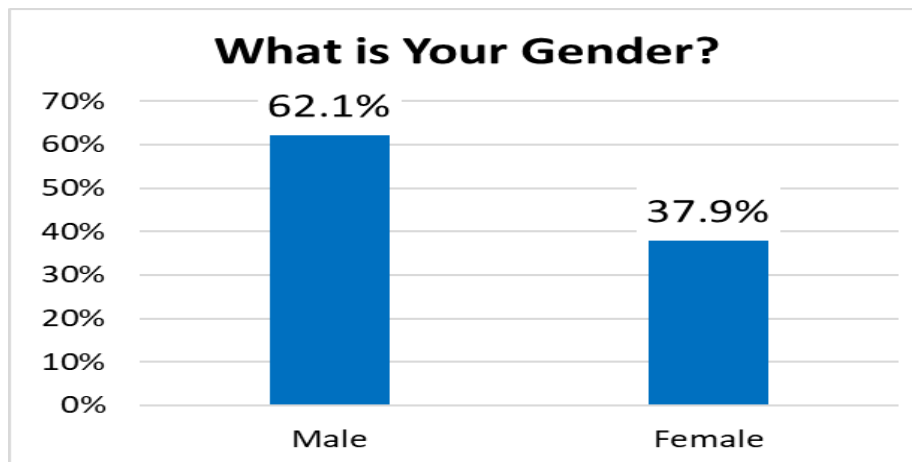
The purpose of having the open-ended questions in addition to the fixed response survey questions was to give the general public an option to voice their opinions which we might not have been able to be expressed through the fixed response questions.

The survey was then rolled out to the general public through the I-80 planning study website. There were a Total of 6312 responses to the survey when it was closed in August 2018.

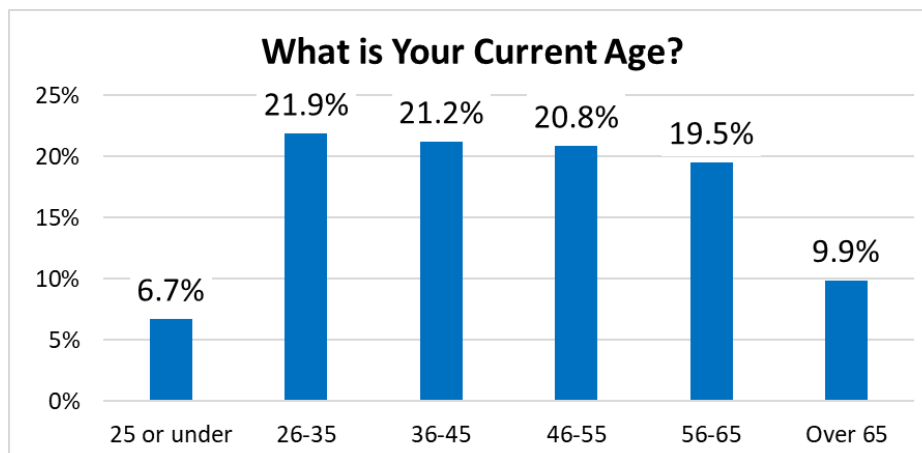
## RESULTS

### Evaluation and Analysis of Quantitate Data

The first few questions of survey were focused on understand the demographics of the public that was responding to the survey. Of the 6312 survey responses, 62.1% of the respondents were male and 37.9% were female. There were only 6.7% of the responses from people aged less than 25 and about 10% from people over 65 years of age. Most of the responses were from people in the age brackets 26-35, 36-45, 46-55 and 56-65 with approximately 20% of responses in each age bracket.

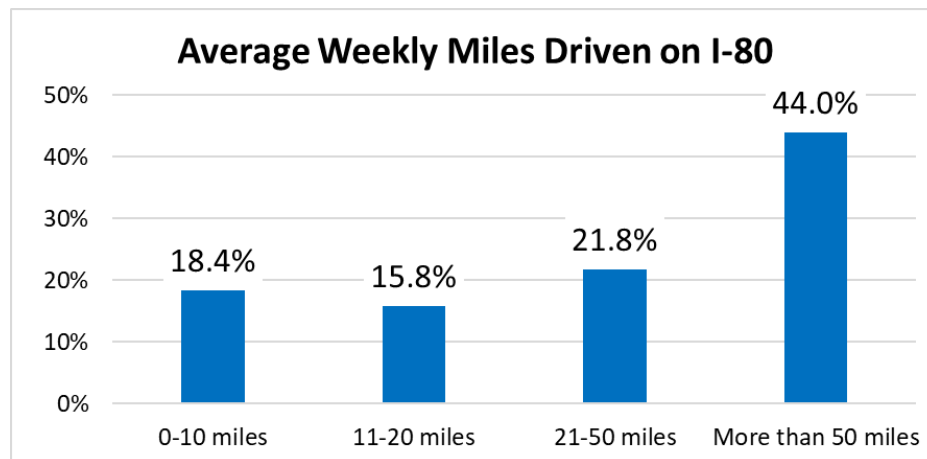


*Figure 1: Gender Distribution*

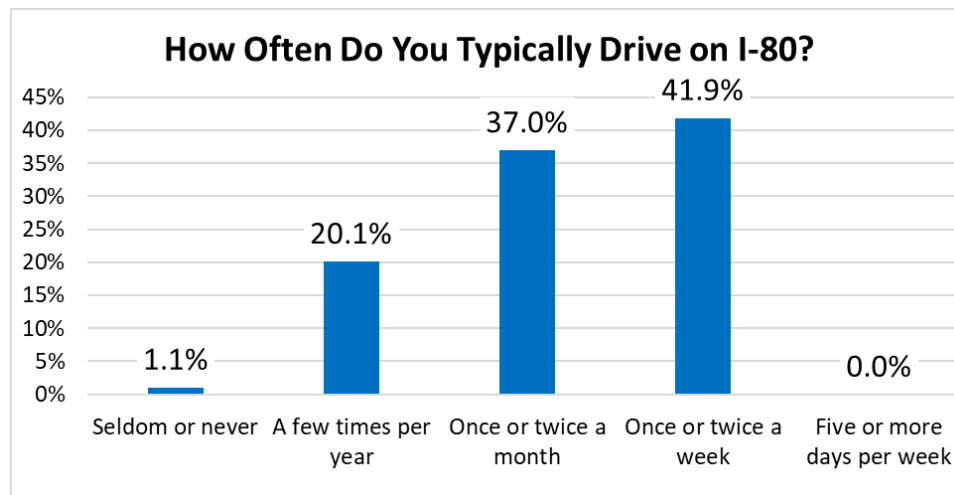


*Figure 2: Age Distribution*

On an average, 44% of the survey respondents travel more than 50 miles on the Interstate every week, 21.8% travel more than 20 and less than 50 miles, 15.8% travel more than 11 miles and less than 20 miles and 18.4% travel less than 10 miles per week. Almost none of the respondents travel five or more times per week on the interstate, 42% of them travel once or twice a week, 37% of them travel once or twice a month, 20% of them travel a few times a year and about 1% of them seldom or rarely travel on the I-80. 47.4% of the respondents use the Interstate for pleasure – weekend or vacation trips, 28.7% of them use the I-80 as a part of their commute to and from work and the remaining 24% drive on the Interstate as a part of their work.

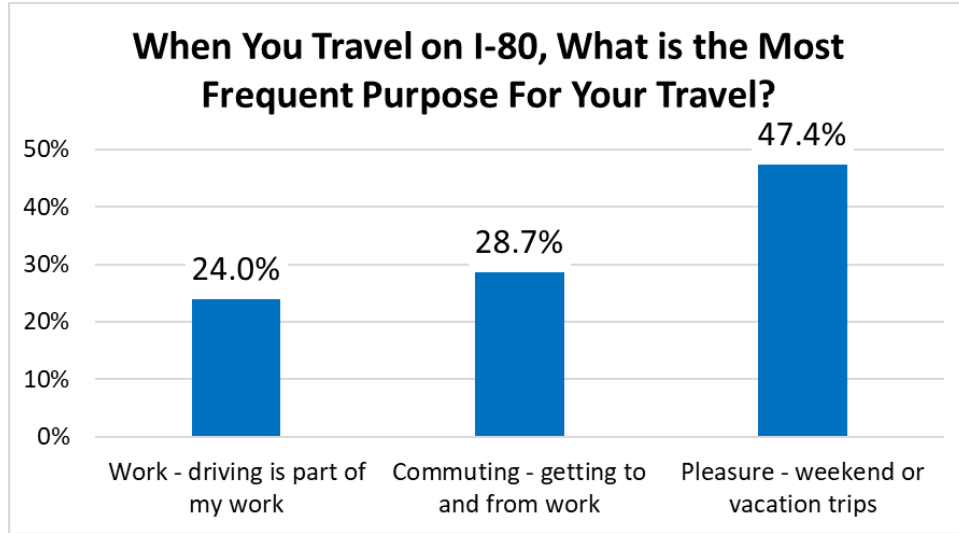


*Figure 3: Average weekly miles driven on I-80*



*Figure 4: Frequency of Travel on I-80*





*Figure 5: Purpose of travel on I-80*

When asked about their opinions on the Level of traffic when they use the I-80, only 20.4% of the people said that the traffic is unacceptably high, 45% percent of the people said that the traffic is high, but it is acceptable, 31.5 said that the traffic is moderate and the remaining 3% said low traffic. We tried to find out the people's opinions about certain aspects and asked them to grade them on a scale of 1 to 5, 1 being dissatisfied and 5 being satisfied. The average ratings of the aspects were:

Flow of traffic - 2.88

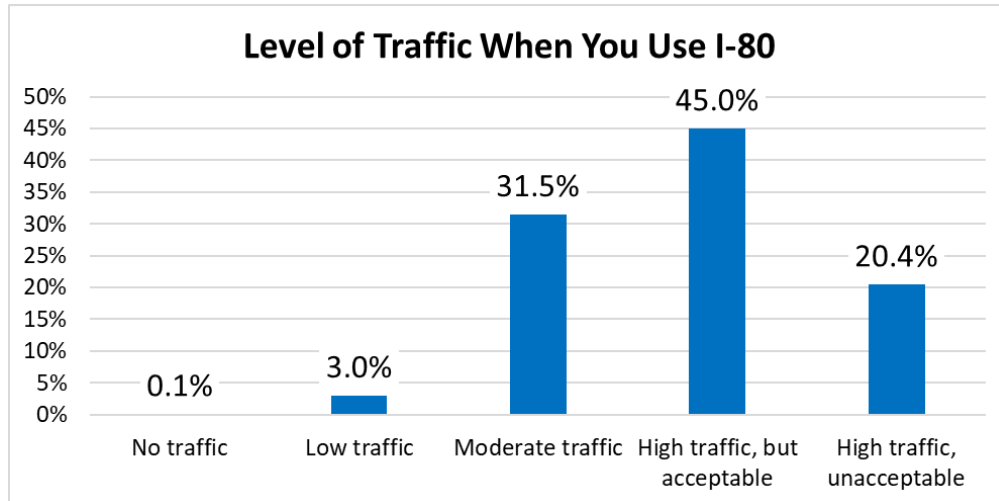
Amount of Truck Traffic – 2.39

Road Condition – 3.39

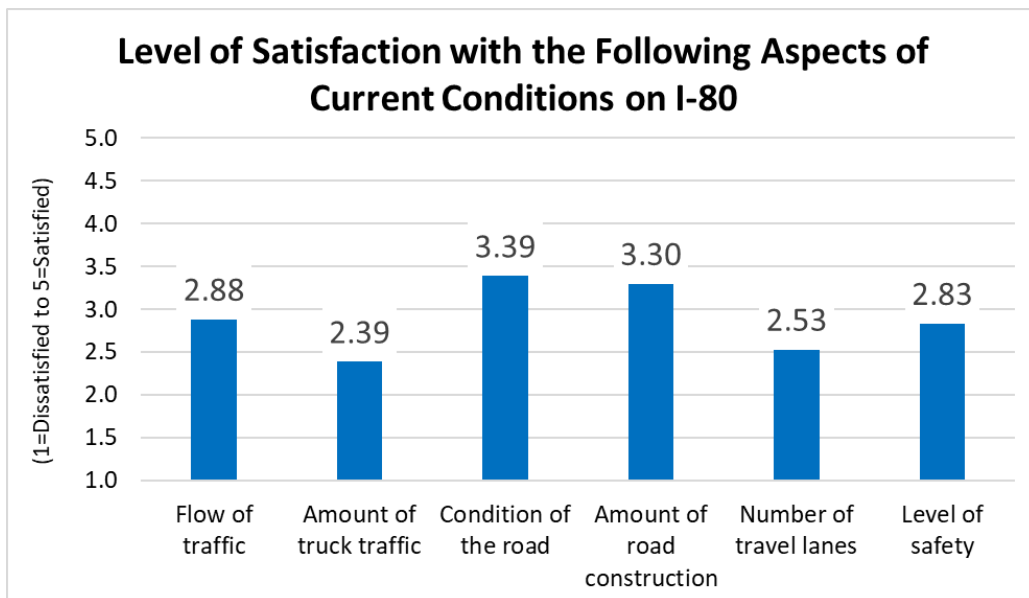
Amount of road construction – 3.30

Number of travel lanes – 2.53

Level of Safety – 2.83

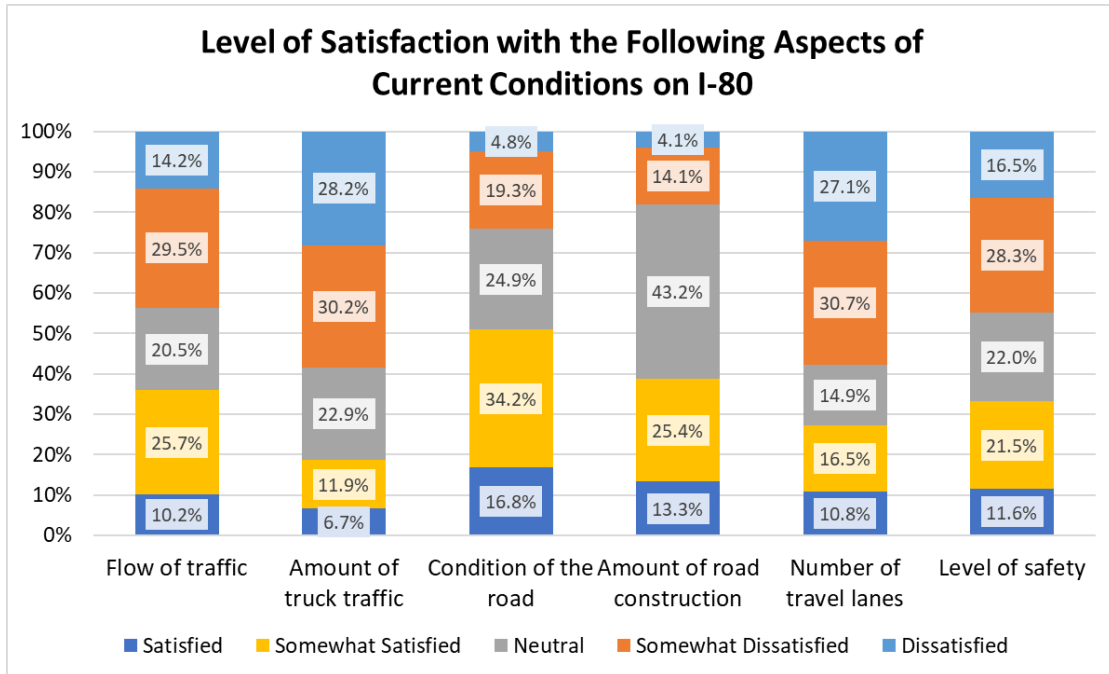


*Figure 6: Level of traffic on I-80*

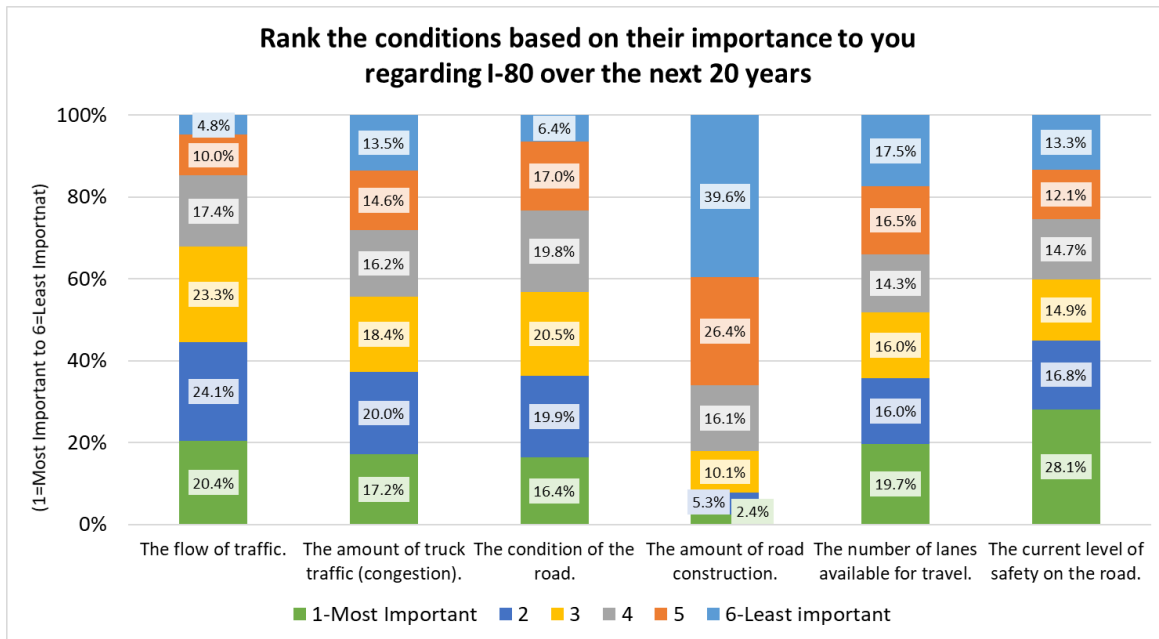


*Figure 7: Current I-80 conditions - 1*

Below are the visualizations of the same aspects mentioned above broken down by percentage of level of satisfaction for each aspect and the importance of each of these aspects of I-80 for the people in the next 20 years:



**Figure 8: Current I-80 conditions - 2**

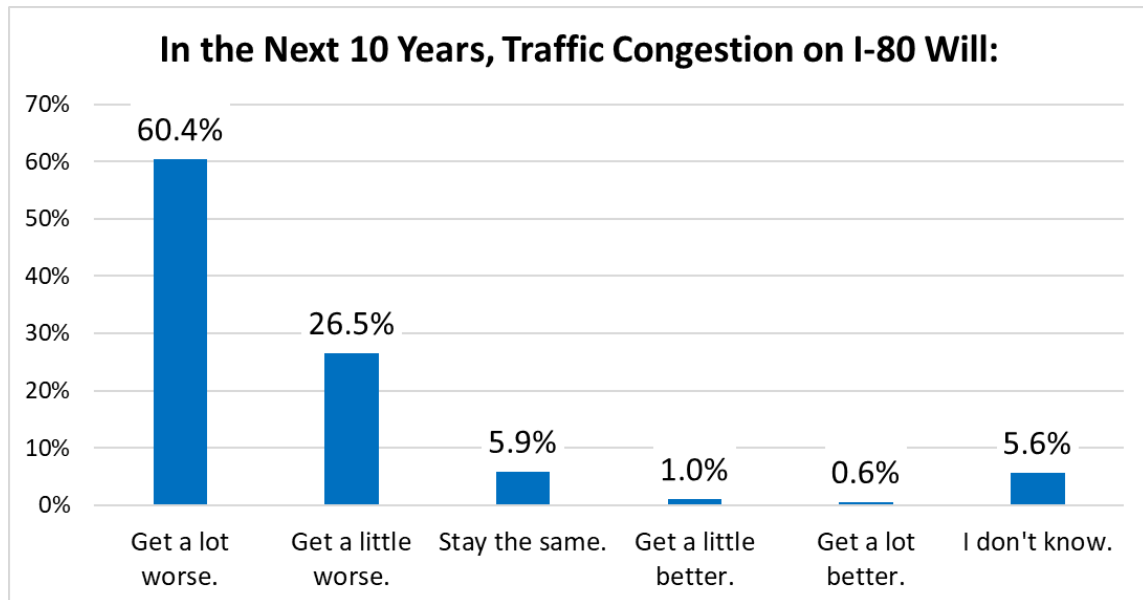


**Figure 9: Importance of future I-80 conditions**

It is interesting to see how people are not bothered by the amount of road construction, if that would mean better infrastructure in the future. Most of the people ranked Flow of traffic and Road Safety with high importance.

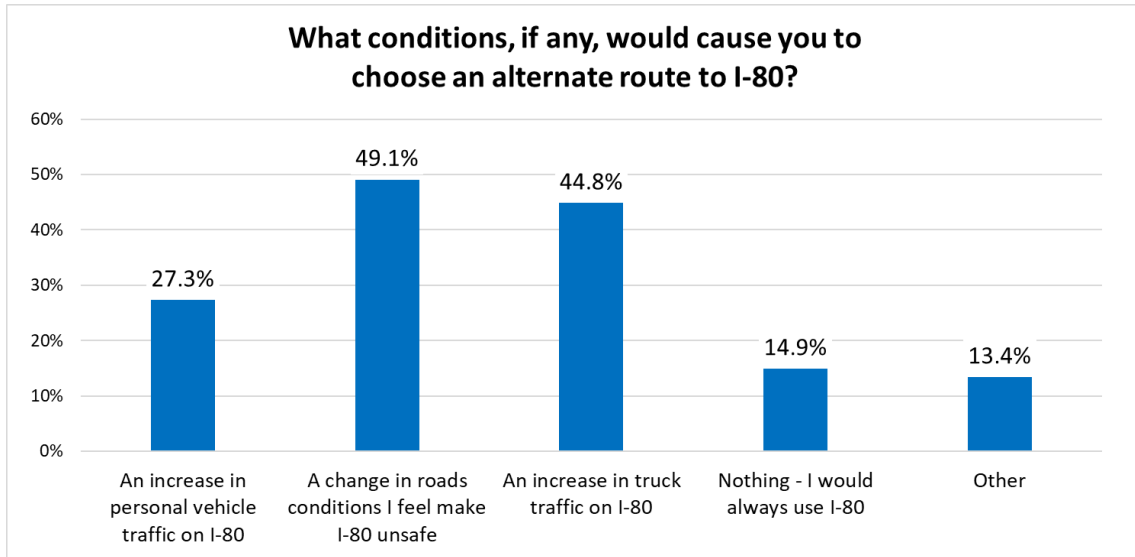
In terms of people's opinion on how the traffic congestions would be on the I-80 in the next 10 years, 60% of them said that it would get a lot worse and 26.5% of said it would get a little worse.

There are less than 2% people who think that the traffic congestion would get better.



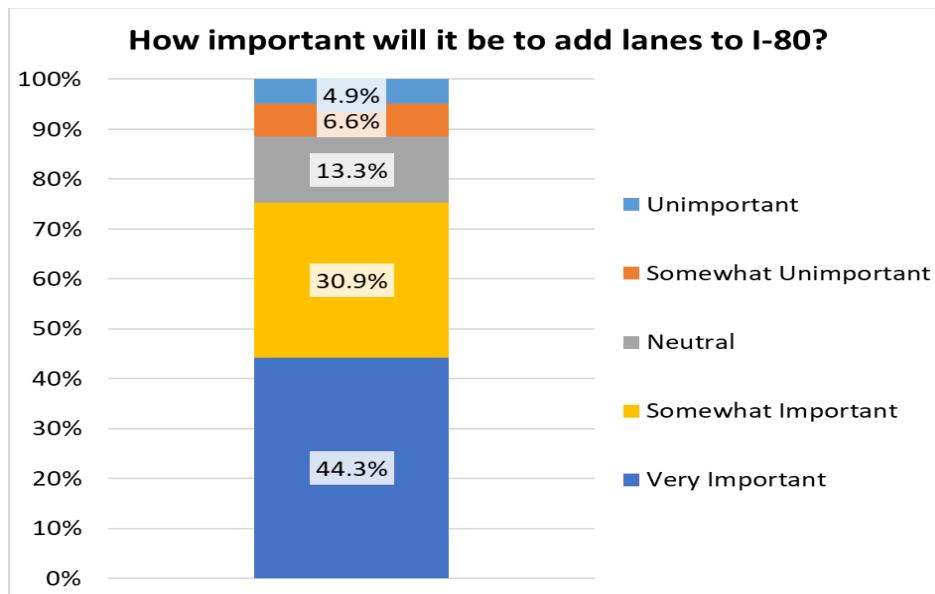
*Figure 10: Traffic Congestion on I-80 in next 10 years*

While trying to learn the reasons because of which people would choose to take an alternative route to I-80, safety again turns out to be the highest concern with 49% of the people saying lesser safety on I-80 would cause them to take an alternative route. 45% of them would choose to take an alternative route if the truck traffic increased and 27% would choose an alternative route if the general traffic would increase. 15% would always continue to use the Interstate and 13% of them had other reasons because of which they would switch routes.



**Figure 11: Condition that would result in taking alternative routes**

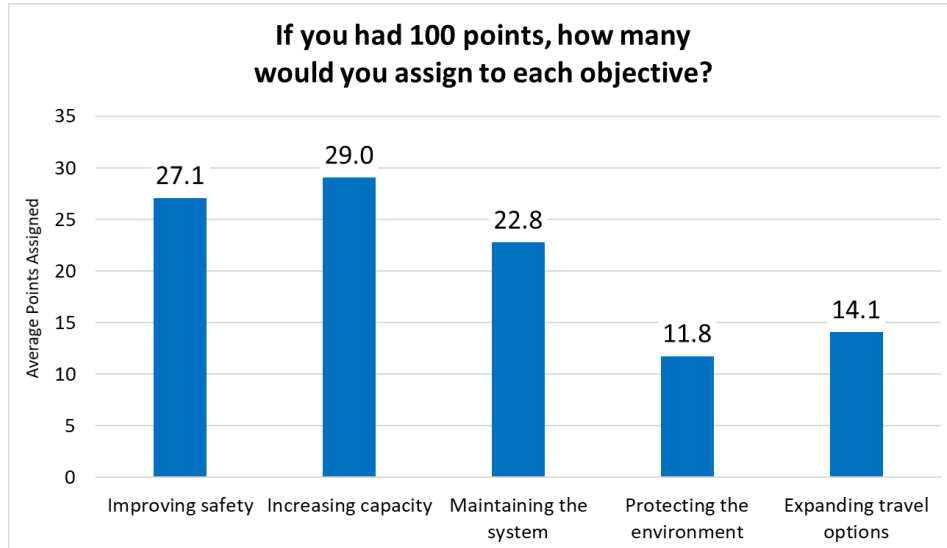
We learned that it is a major portion of the people think it is very important/ important to add lanes to I-80. 44.3% think that it is very important and 30.9% of them think that it is somewhat important. 13% had neutral opinion on this and the rest thought it was unimportant.



**Figure 12: Importance of adding lanes to I-80**

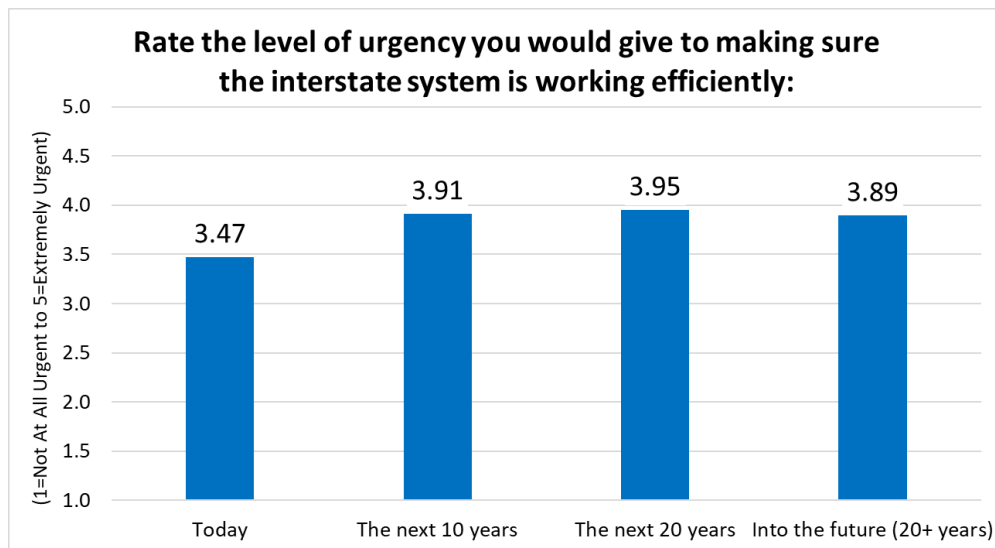
To learn the people’s opinions on some of the factors to consider during the development activities, we asked them to assign points to each of these categories depending on their importance giving

them a total points of 100. On an average, Increasing capacity and improving safety received the highest points with 29 and 27.1 average points each. Protecting the environment received the least points with only 11.8 average points.



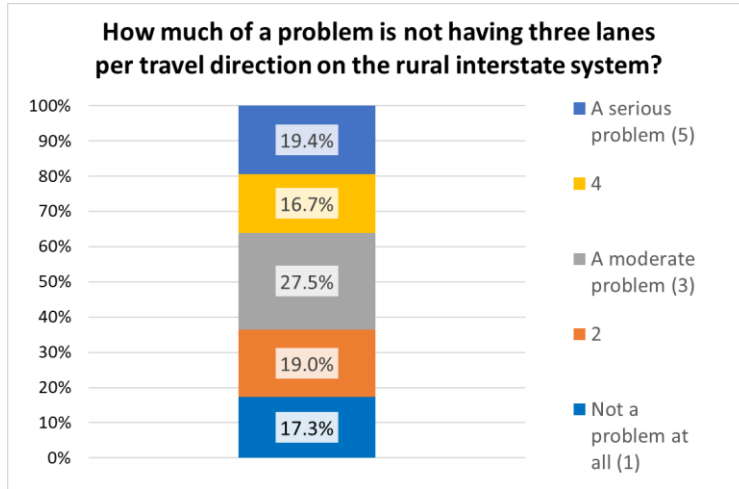
*Figure 13: People's opinions on various factors*

The bar chart below shows the level of urgency that the people think the interstate system should be working effectively at which time frames and all time frames which are Today, the next 10 years, the next 20 years and 20+ years received almost the same level of urgency grading.



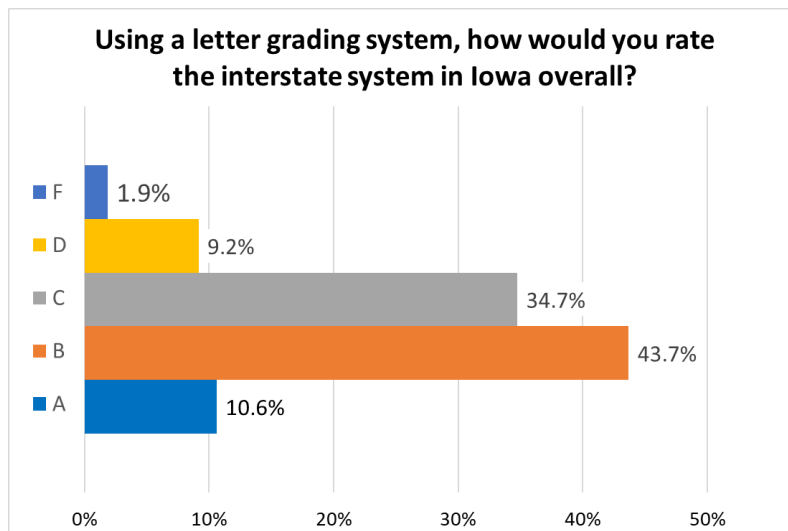
*Figure 14: Level of Urgency*

Most people think that not having three lanes per travel direction on the rural interstate system is a moderate problem with almost equal percentage of the remaining people thinking that it is a problem and it is not a problem.



**Figure 15: Importance of having three lanes on the rural interstate system**

Overall, on a letter grading system, 43.7 percent of the people rated the interstate system in Iowa as B, 34.7 percent rated it as C and 10.6% rated it as A. 9% rated it as D and about 2% of them rated it as F.



**Figure 16: Overall Interstate System rating**

### **Evaluation and Analysis of Qualitative Data**

The last 2 questions of the survey had open ended questions and people could express their views in free text about changes or improvements to interstate that would have positive and negative impacts on their lives in two different questions.

### **Data Preparation and algorithm execution**

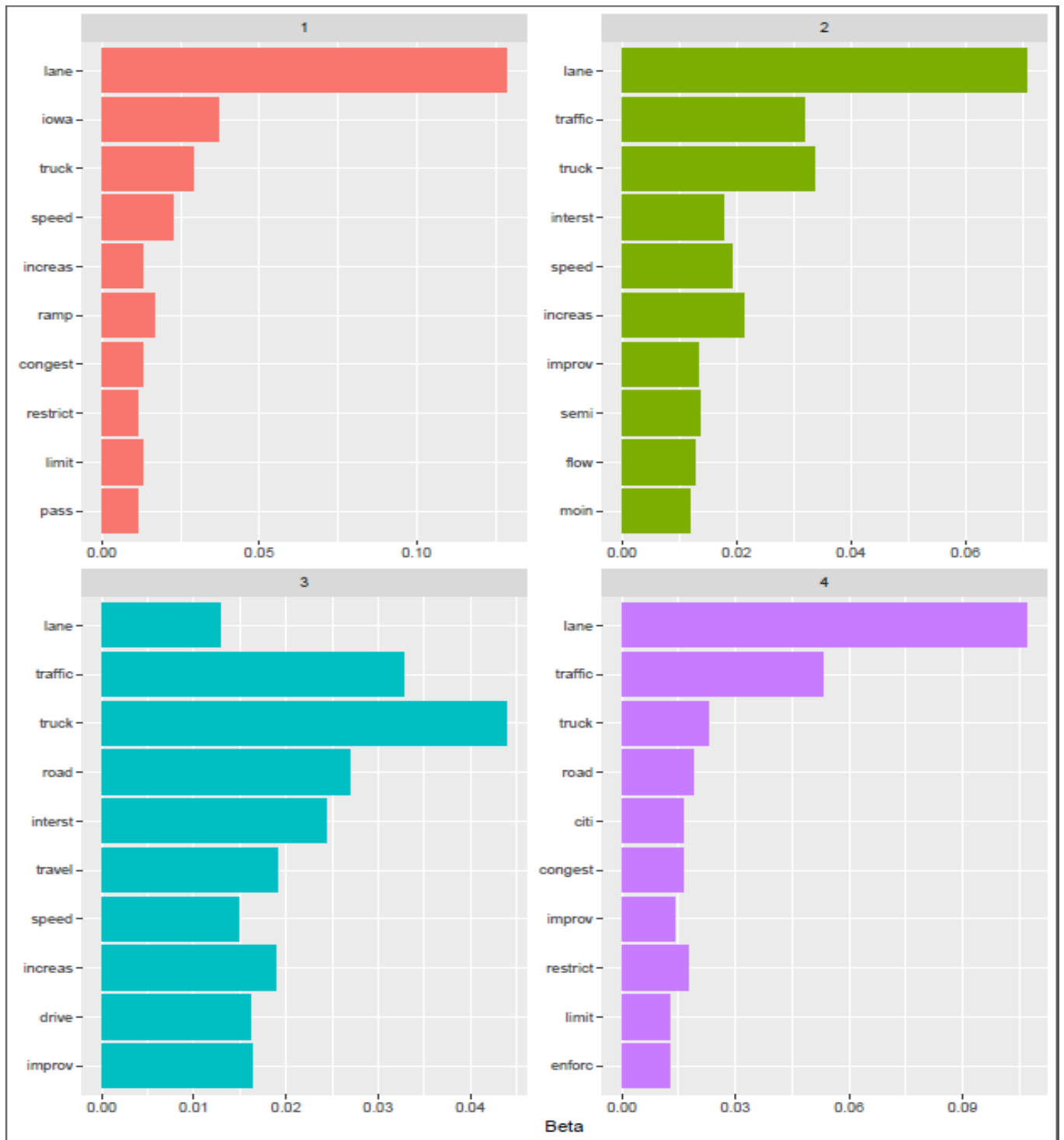
The data was stored in an excel file which was imported into RStudio. The two columns were then split into 2 separate variables which would make it easier to run the algorithm to extract themes about the positive impacting factors and negative impacting factors separately. The free text questions were not mandatory fields. So, there were a lot of blanks responses that needed to be removed for the analysis. Upon eliminating the blanks, we had 3718 responses in each of the columns. Since we will have to run the algorithm multiple times making tweaks each time, we decided to define a function in R which would have all the preprocessing and algorithm execution steps defined into it. This way, we could run it multiple times by just changing the variables in the function. We then defined the function to create a corpus object, convert it to a document term matrix and use it as an input for the LDA algorithm. We then plotted the top terms in each of the themes extracted by the LDA algorithm as a bar plot with the words on the Y Axis and “Beta” on the X axis which is the probability of the word belonging to that specific theme. Naturally, when we successfully ran the code first time, there were a lot of general English vocabulary that wouldn't really have any information when analyzed as individual words. So, we removed the stop words in the text and ran the algorithm again. Glancing at the results, we could see words which are fundamentally same but structured differently, i.e. derived from the same base word (E.g. increase, increasing, increased and also spelling mistakes like increasd). So, we stemmed the words in the text which would trim all these words to their base “increas” and all of these would be considered as a single term.



After all the above preprocessing and algorithm building, we ran the algorithm for both positive and negative impacting factors and we could see some pattern developing but the themes were not discrete. There were words like “lane” which were appearing in all the themes with high probability and it looked like this certain word might be skewing the results and making it difficult for the algorithm to extract discrete themes. There were also words like “iowa”, “des” “moines” etc. which were names of places and not adding much value to the analysis. So, we removed the word “lane” and other custom stop words from the text and ran the algorithm again. The results looked much better than the previous result, but we were still not able to derive discrete themes as we would have liked from them. So, we decided to check if bigrams and trigrams would be able to give us better information compared to single words. Although not very discrete, we could see themes like “Lane Management”. “speed control”, “truck traffic management” emerging from the results at this point. We even tried increasing and decreasing the number of themes and still the results were not conclusive. While looking at the bigrams and trigrams, it looked like they held much more information by themselves rather than being used in the LDA algorithm. So, we decided to just plot the bigram and trigram frequencies without any algorithms and get a feel of what most people were trying to express through free text.

The results of each above-mentioned steps are displayed below for each of the questions:

**1. What changes or improvements to interstate transportation would impact you/your life in a positive way?**



*Figure 17: Preliminary LDA Results - Positive Impact*

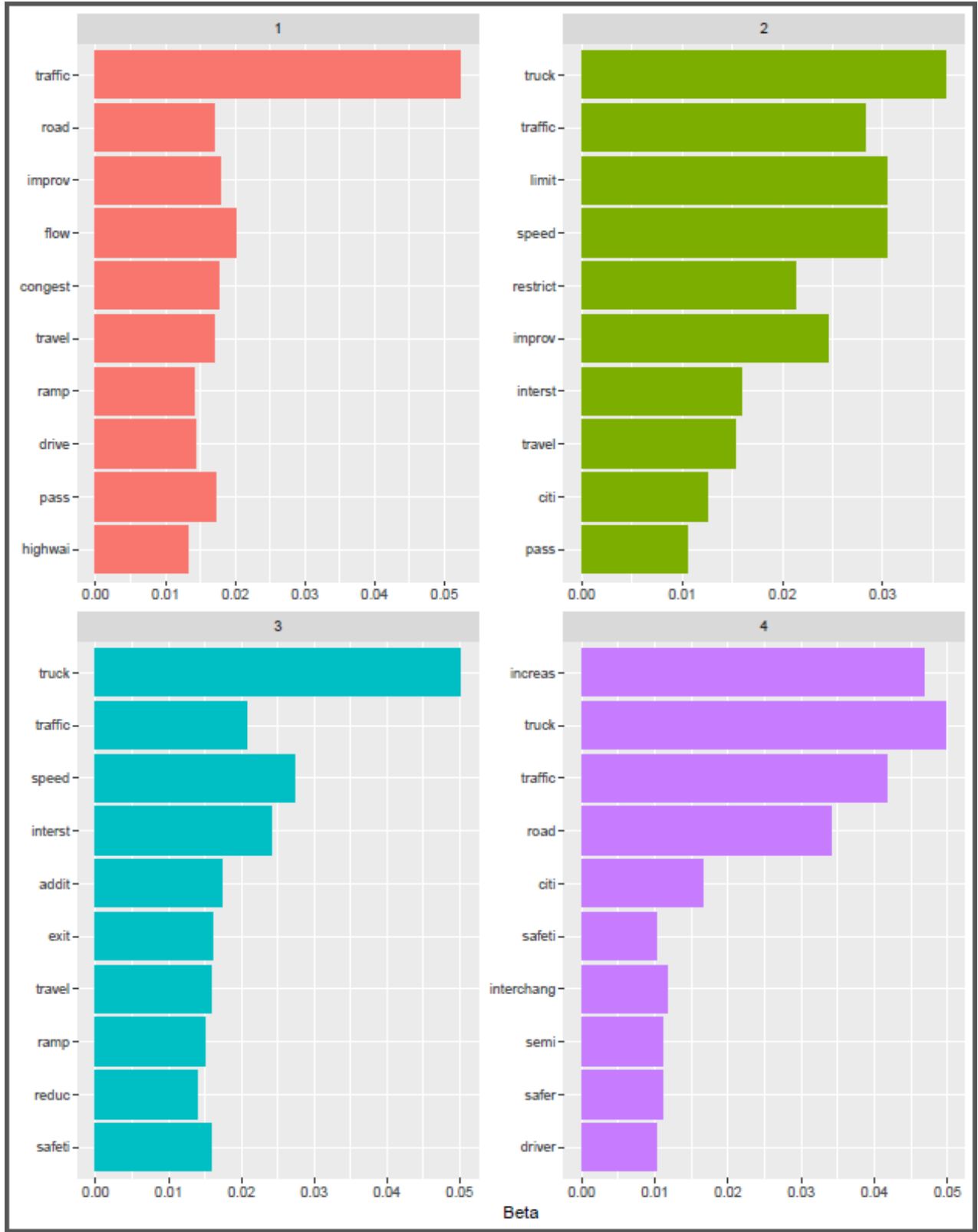


Figure 18: LDA results -without custom words - Positive Impact

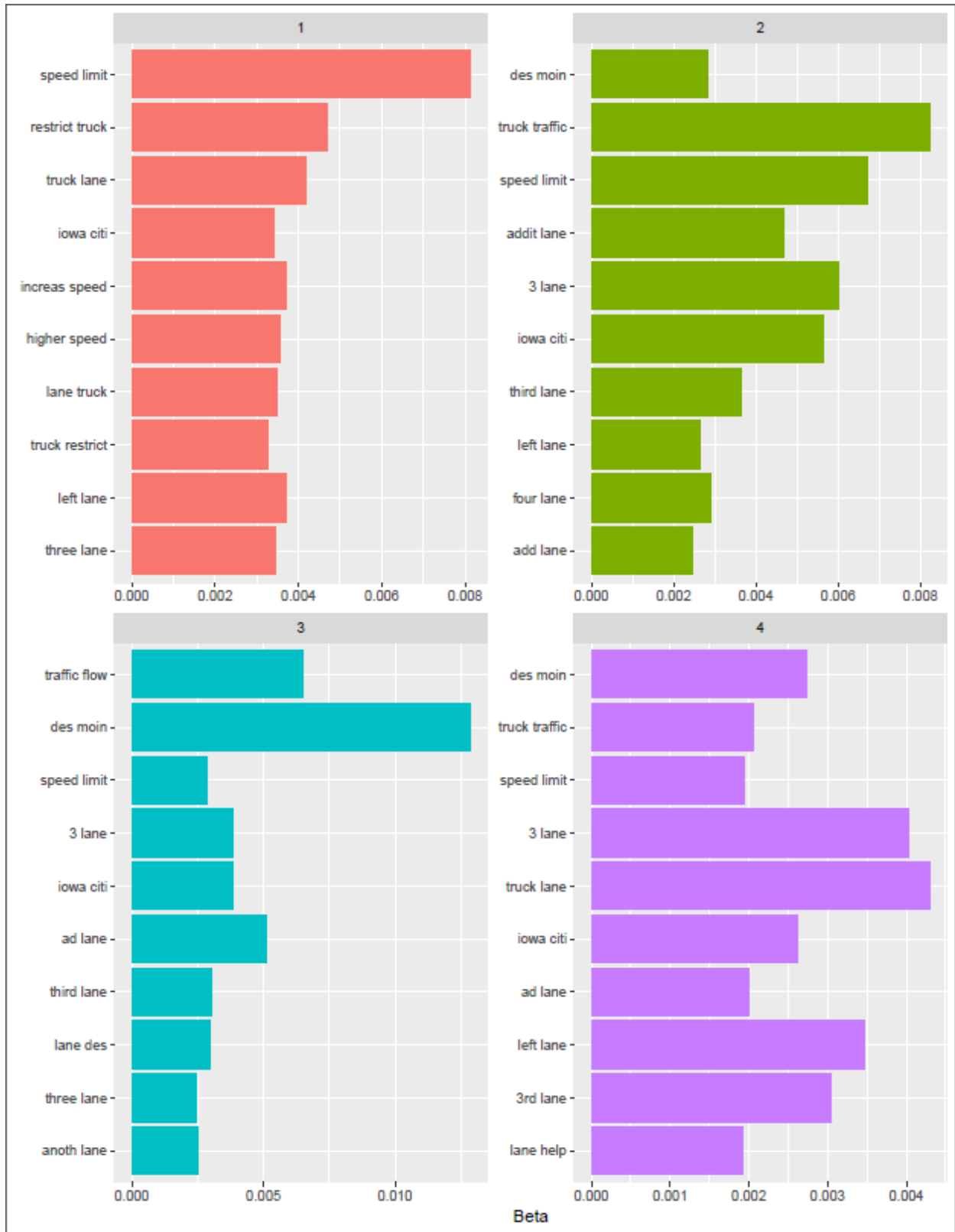
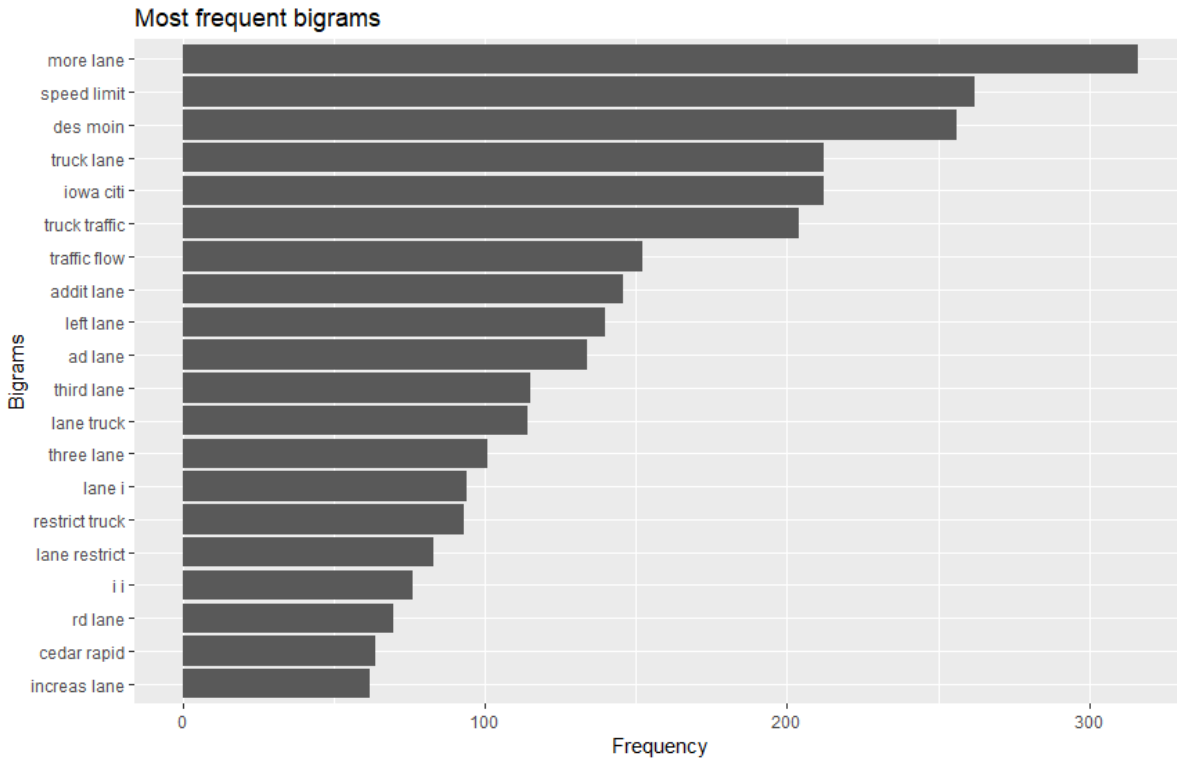


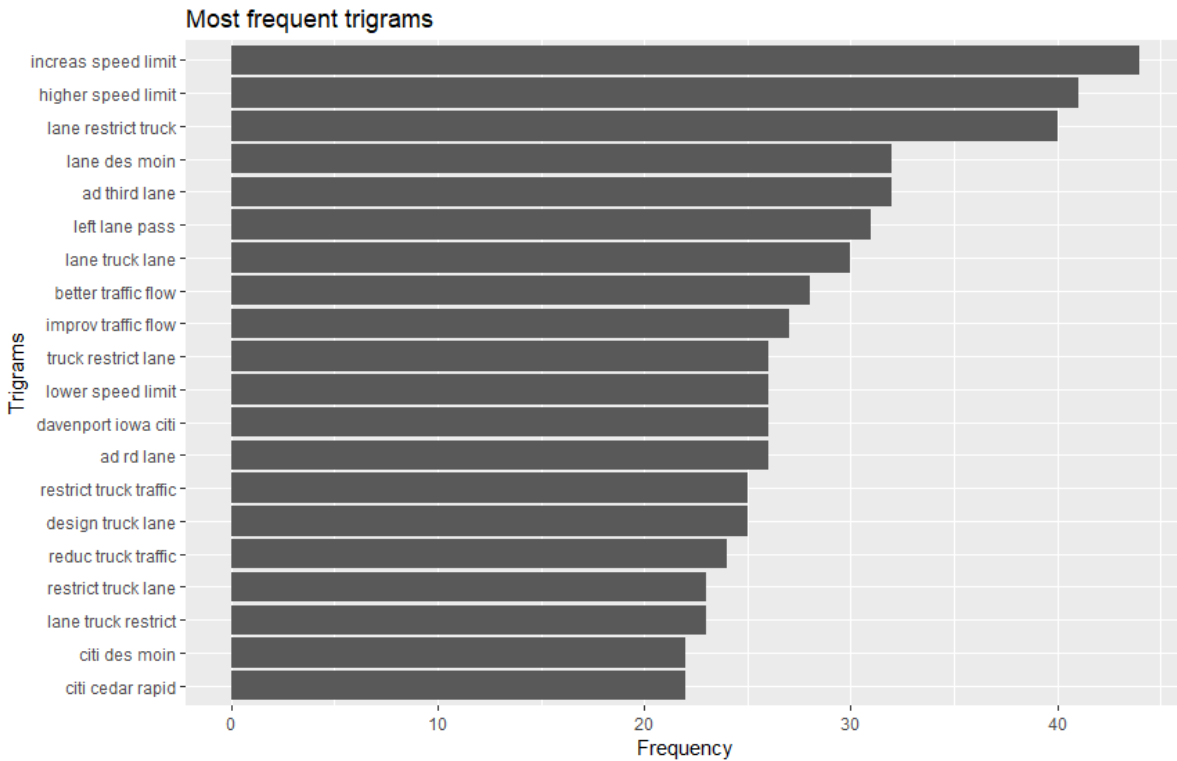
Figure 19: LDA Results - Bigrams - Positive Impact



Figure 20: LDA results - Trigrams - Positive Impact



**Figure 21: Bigram Frequencies - Positive Impact**



**Figure 22: Trigram frequencies - Positive Impact**

2. What changes or improvements to interstate transportation would impact you/your life in a negative way?



Figure 23: Preliminary LDA results - Negative Impact

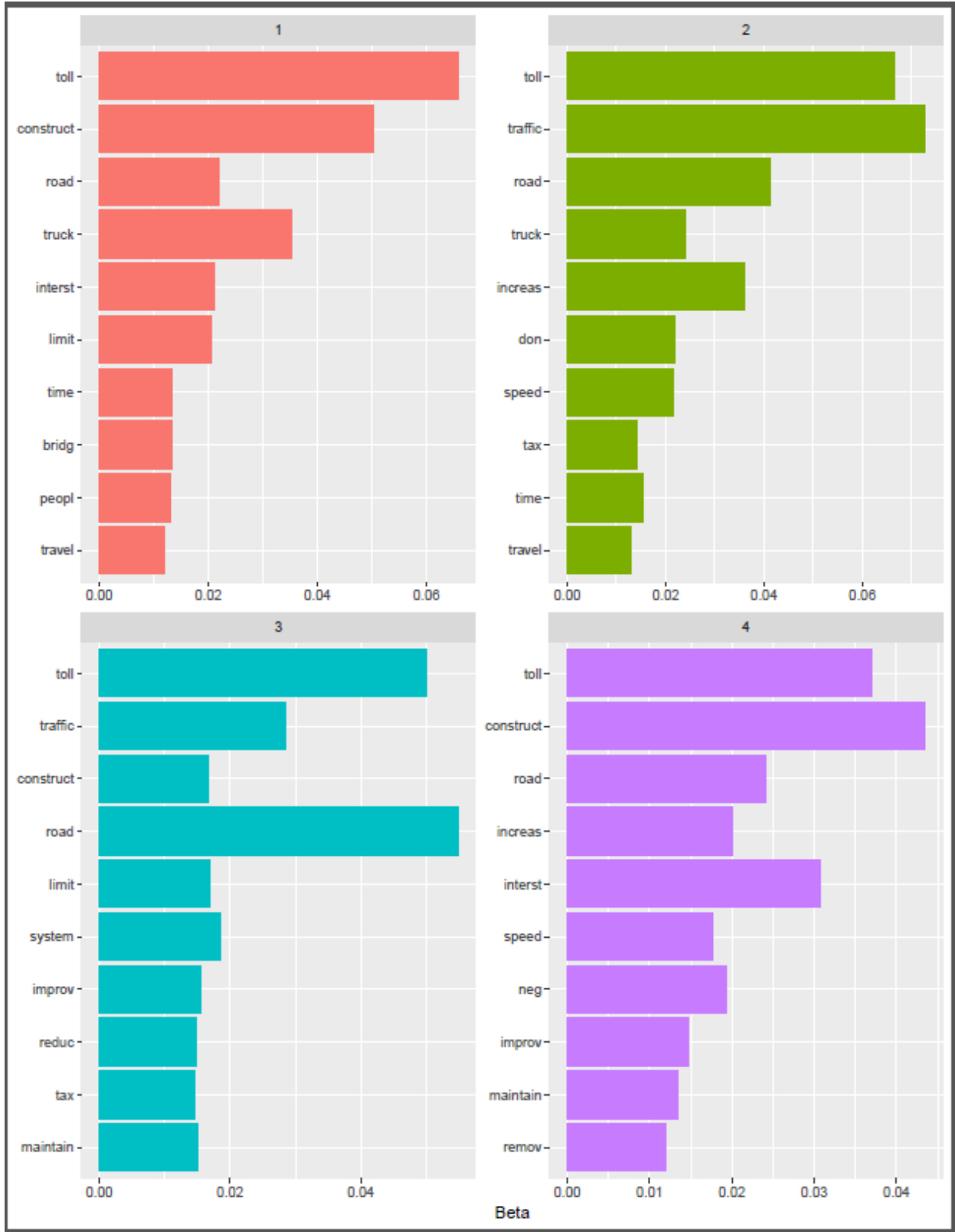


Figure 24: LDA results - without custom words - Negative Impact



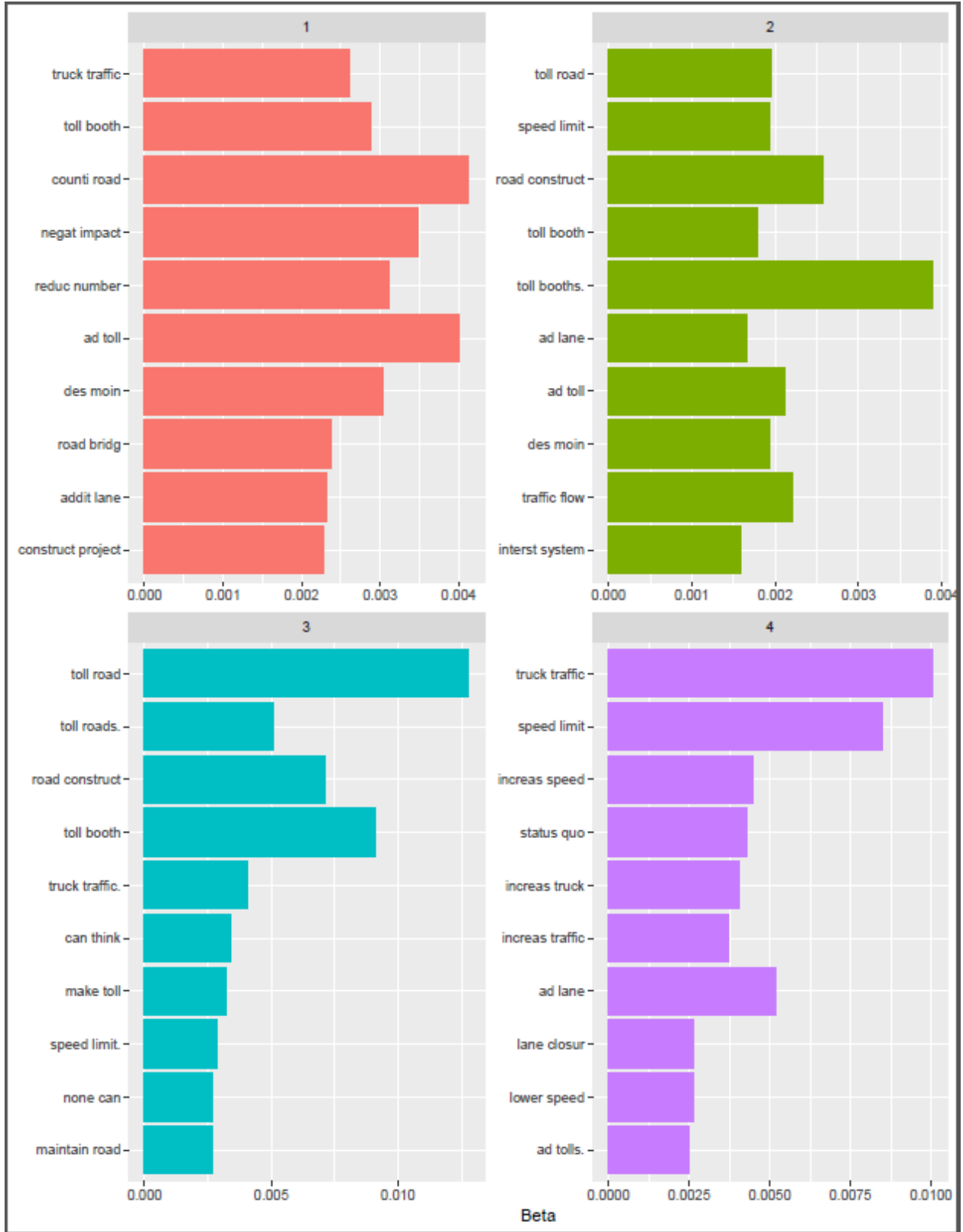


Figure 25: LDA results - Bigrams - Negative Impact



Figure 26: LDA Results - Trigrams - Negative Impact

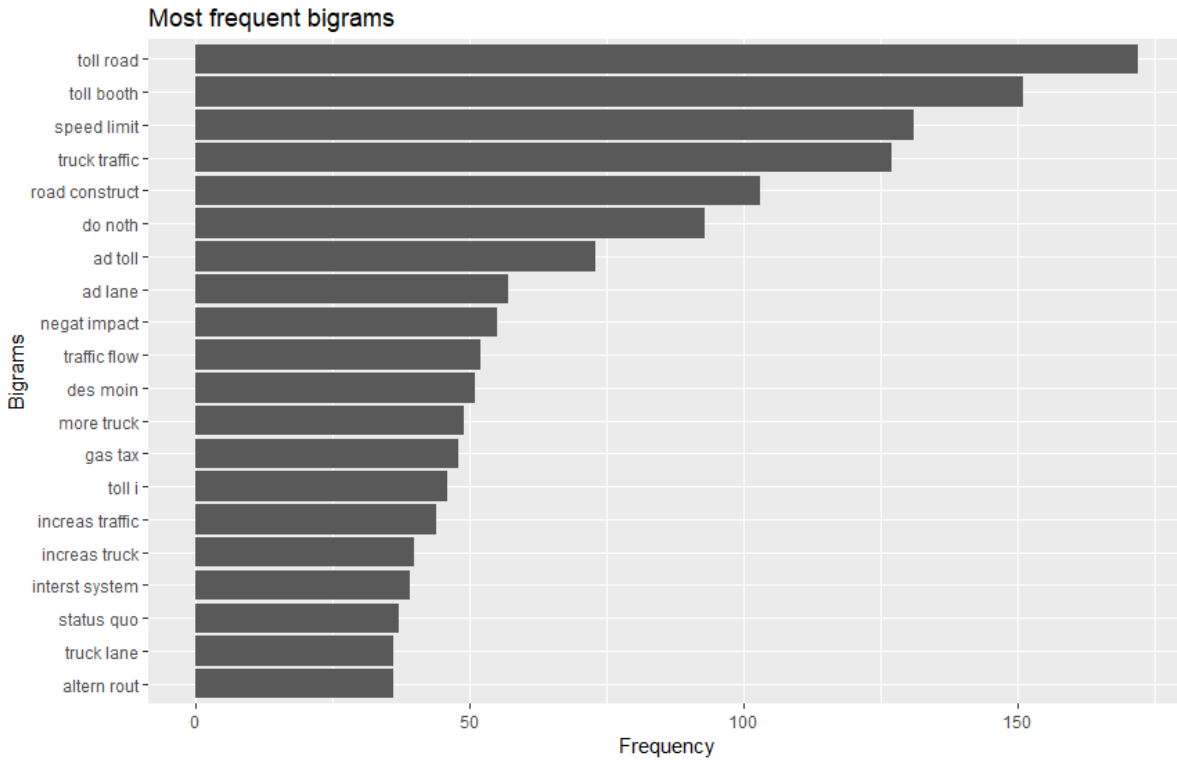


Figure 27: Bigram Frequencies - Negative Impact

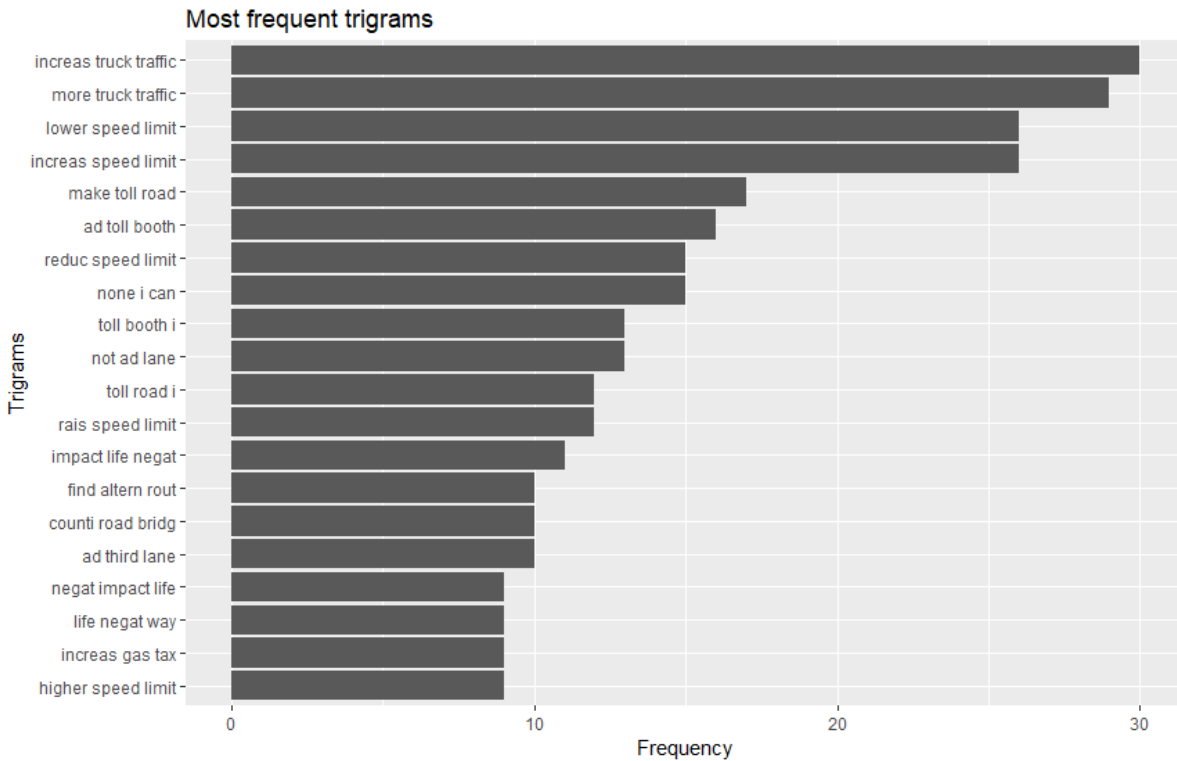


Figure 28: Trigram Frequencies - Negative Impact

## LIMITATIONS

LDA, like every other algorithm is not perfect. It has its limitations and disadvantages. Some of them are:

- LDA does not work well if the design is not balanced (i.e. the number of objects in various classes are (highly) different).
- LDA is sensitive to overfit and validation of LDA models is at least problematic.
- LDA is not applicable (inferior) for non-linear problems
- Small Sample Size problem

Some of these limitations of LDA could give us more insights into why the algorithm didn't work the best in our case. Although 6312 survey responses are a lot, seems like it is still small for the LDA Algorithm to work optimally. Manually skimming through some of the survey responses, it could be observed that a really large section of the people's responses were really small (3 – 5 words) and most of the customers talked about the same major issues like additional lanes, no toll booths, truck traffic etc. which could have caused a decrease in the overall vocabulary that is available for analysis. Sometimes all these different issues were mentioned in a single response which might have made it difficult for the algorithm to draw discrete patterns from the text. All these factors might have caused LDA, which otherwise is a good method for topic modelling, to not work as optimally as it usually can.

## CONCLUSION

As a part of the Interstate-80 Study, the Iowa DOT management now has answers to the various questions that they were seeking public opinion on. At a glance, they can see the demographics of the survey respondents, their opinions on the current conditions, future conditions and various other factors and their importance. The bigram and trigram frequencies at the end of the analysis give a picture of the kind of changes that would have positive and negative impacts on the people's lives. While talking about the factors that would affect them positively, people were talking about various things like increasing speed limit/higher speed limit, restricted truck lanes, additional lanes, better traffic flow etc. While talking about factors that would affect them negatively, people were talking about things like adding tolls, increased truck traffic, doing nothing (i.e. leaving the Interstate in its current condition), raising gas tax, road construction etc. The results of this analysis will definitely go a long way in factoring public opinions while making decisions of development and investment plans for the Interstate.

**REFERENCES**

- [1] Berry, M.B. and Koqan, J. (2010) Text Mining: Applications and Theory. st 1 edn. Colorado: Wiley.
- [2] Srivastava, A. and Sahami, M (2009) Text Mining: Classification, Clustering, and Applications. st 1 edn. USA: Chapman and Hall/CRC.
- [3] Berry, M.W. (2003) Survey of Text Mining: Clustering, Classification, and Retrieval. nd 2 edn. New York: Springer Publishing.
- [4] Franke, J., Nakhaeizadeh, G. and Renz, I. (2003) Text Mining: Theoretical Aspects and Applications. st 1 edn. Berlin: Physica-Verlag HD.
- [5] BLEI, D. M. & JORDAN, M. I. Modeling annotated data. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003. ACM, 127-134
- [6] Hamed Jelodar , Yongli Wang , Chi Yuan , Xia Feng, Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey
- [7] Rubayyi Alghamdi, Khalid Alfalqi, A Survey of Topic Modeling in Text Mining, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 1, 2015
- [8] <https://iowadot.gov/interstatestudy/home>

**APPENDIX A – R Code****R code for generating wordclouds:**

```
library(tm)

library(SnowballC)

library(wordcloud)

## Setting working directory and reading data

setwd("W:/PerformanceTechnology/AssetMgmt/I-80 Corridor Study Survey")

data <- read.csv("data.csv")

# reading Column 19 and Column 20 separately to analyze positive and negative
comments separately

pos <- data$Positive

neg <- data$Negative

# replacing blanks with "NA"

pos[pos==""] <- NA

neg[neg==""] <- NA

#checking the number of NA's

sum(is.na(pos))

sum(is.na(neg))

# removing NA's from the columns

pos <- na.omit(pos)

neg <- na.omit(neg)

# creating a variable to easily switch between positive and negative analysis
```

```

X <- neg

# function to create the word cloud

y <- function(x)
{
  corpus <- Corpus(VectorSource(x))

  corpus <- tm_map(corpus, PlainTextDocument)

  corpus <- tm_map(tm_map(tm_map(corpus, stripWhitespace), tolower),
stemDocument)

  corpus <- tm_map(corpus, removeWords , stopwords("english"))
  corpus <- tm_map(corpus, removePunctuation)
  corpus <- Corpus(VectorSource(corpus))

  wordcloud(corpus, max.words = 100, random.order = FALSE,
colors=brewer.pal(8, "Dark2"))
}

y(neg)

```

**R code for the entire process including setting up the environment, data preparation and plotting the results:**

```

#setting up the environment

install.packages("tidyverse")

library(tidyverse)

install.packages("tidytext")

```



```

library(tidytext)

install.packages("topicmodels")

library(topicmodels)

install.packages("tm")

library(tm)

install.packages("SnowballC")

library(SnowballC)

setwd("W:/PerformanceTechnology/AssetMgmt/I-80 Corridor Study Survey")

data <- read.csv("data.csv")

pos <- data$Positive

neg <- data$Negative

pos[pos==""] <- NA

neg[neg==""] <- NA

sum(is.na(pos))

sum(is.na(neg))

pos <- na.omit(pos)

neg <- na.omit(neg)

# function to get & plot the most informative terms by a specified number
# of topics, using LDA

LDA_topterms <- function (pos, # data
                          plot = T,
                          number_of_topics = 4)
{

```

```

# create a corpus and document term matrix

Corpus <- Corpus(VectorSource(pos)) # making a corpus object

DTM <- DocumentTermMatrix(Corpus) # get the count of words/document

# removing empty terms in the DTM

unique_indexes <- unique(DTM$i) # get the index of each unique value

DTM <- DTM[unique_indexes,] # get a subset of only those indexes

# preform LDA & get the words/topic in a tidy text format

lda <- LDA(DTM, k = number_of_topics, control = NULL)

topics <- tidy(lda, matrix = "beta") # convert the LDA output to a tidy

# get the top ten terms for each topic

top_terms <- topics %>% # take the topics data frame and..

  group_by(topic) %>% # treat each topic as a different group

  top_n(10, beta) %>% # get the top 10 most informative words

  ungroup() %>% # ungroup

  arrange(topic, -beta) # arrange words in descending informativeness

# if the user asks for a plot (TRUE by default)

if(plot == T){

  # plot the top ten terms for each topic in order

  top_terms %>% # take the top terms

    mutate(term = reorder(term, beta)) %>% # sort terms by beta

```

```

ggplot(aes(term, beta, fill = factor(topic))) + # plot beta by theme
geom_col(show.legend = FALSE) + # as a bar plot
facet_wrap(~ topic, scales = "free") + # which each topic in a separate plot
labs(x = NULL, y = "Beta") + # no x label, change y label
coord_flip() # turn bars sideways
}else{
# if the user does not request a plot
# return a list of sorted terms instead
return(top_terms)
}
}

```

```

LDA_topterms(pos,T,2)
# remove stopwords
newsDTM <- pos %>%
  VectorSource() %>%
  Corpus() %>%
  DocumentTermMatrix() %>% # create a document term matrix to clean
  tidy() %>% # convert the document term matrix to a tidytext corpus
  anti_join(stop_words, by = c("term" = "word"))
head(newsDTM)
LDA_topterms (newsDTM$term)

```

```
#stemming

newsDTM1 <- newsDTM %>%

  mutate(stem = wordStem(term))

LDA_topterms (newsDTM1$stem, F, 6)
```

### **R code for running the algorithm without custom stop words**

```
#setting up the environment

install.packages("tidyverse")

library(tidyverse)

install.packages("tidytext")

library(tidytext)

install.packages("topicmodels")

library(topicmodels)

install.packages("tm")

library(tm)

install.packages("SnowballC")

library(SnowballC)

setwd("W:/PerformanceTechnology/AssetMgmt/I-80 Corridor Study Survey")

data <- read.csv("data.csv")

pos <- data$Positive

neg <- data$Negative

pos[pos==""] <- NA
```

```

neg[neg==""] <- NA

sum(is.na(pos))

sum(is.na(neg))

pos <- na.omit(pos)

neg <- na.omit(neg)

pos <- neg

# function to get & plot the most informative terms by a specified number
# of topics, using LDA

top_terms_by_topic_LDA <- function(pos, # should be a column from a dataframe
                                   plot = T, # return a plot? TRUE by default
                                   number_of_topics = 4) # number of topics (4 by default)
{
  # create a corpus (type of object expected by tm) and document term matrix
  Corpus <- Corpus(VectorSource(pos)) # make a corpus object
  Corpus <- tm_map(Corpus, removeWords, c("lane", "lanes", "iowa", "des",
"moines", "moin"))

  DTM <- DocumentTermMatrix(Corpus) # get the count of words/document
  # remove any empty rows in our document term matrix (if there are any
  # we'll get an error when we try to run our LDA)

  unique_indexes <- unique(DTM$i) # get the index of each unique value
  DTM <- DTM[unique_indexes,] # get a subset of only those indexes

  # perform LDA & get the words/topic in a tidy text format
  lda <- LDA(DTM, k = number_of_topics, control = NULL)

```

```

topics <- tidy(lda, matrix = "beta") # convert the LDA output to a tidy
# get the top ten terms for each topic

top_terms <- topics %>% # take the topics data frame and..
  group_by(topic) %>% # treat each topic as a different group
  top_n(10, beta) %>% # get the top 10 most informative words
  ungroup() %>% # ungroup
  arrange(topic, -beta) # arrange words in descending informativeness
# if the user asks for a plot (TRUE by default)
if(plot == T){
  # plot the top ten terms for each topic in order
  top_terms %>% # take the top terms
    mutate(term = reorder(term, beta)) %>% # sort terms by beta value
    ggplot(aes(term, beta, fill = factor(topic))) + # plot beta by theme
    geom_col(show.legend = FALSE) + # as a bar plot
    facet_wrap(~ topic, scales = "free") + # which each topic in a separate plot
    labs(x = NULL, y = "Beta") + # no x label, change y label
    coord_flip() # turn bars sideways
}else{
  # if the user does not request a plot
  # return a list of sorted terms instead
  return(top_terms)
}
}

```

```

top_terms_by_topic_LDA(pos,T,2)

# remove stopwords

newsDTM <- pos %>%

  VectorSource() %>%

  Corpus() %>%

  DocumentTermMatrix() %>% # create a document term matrix to clean

  tidy() %>% # convert the document term matrix to a tidytext corpus

  anti_join(stop_words, by = c("term" = "word"))

head(newsDTM)

top_terms_by_topic_LDA(newsDTM$term)

#stemming

newsDTM1 <- newsDTM %>%

  mutate(stem = wordStem(term))

top_terms_by_topic_LDA(newsDTM1$stem, T, 4)

```

**R code for the entire process including setting up the environment, data preparation and plotting the results – for n-grams (here bigrams and trigrams):**

```

library(tm)

library(SnowballC)

library(topicmodels)

#library(RWeka)

library(tidytext)

```

```

setwd("W:/PerformanceTechnology/AssetMgmt/I-80 Corridor Study Survey")
data <- read.csv("data.csv")
pos <- data$Positive
neg <- data$Negative
pos[pos==""] <- NA
neg[neg==""] <- NA
sum(is.na(pos))
sum(is.na(neg))
pos <- na.omit(pos)
neg <- na.omit(neg)
X <- neg
top_terms_by_topic_LDA <- function(X, # should be a column from a dataframe
                                plot = T, # return a plot? TRUE by default
                                number_of_topics = 4)
{
  corpus <- VCorpus(VectorSource(X))
  corpus <- tm_map(tm_map(tm_map(corpus, stripWhitespace), tolower),
stemDocument)
  corpus <- tm_map(corpus, removeWords , stopwords("english"))
  corpus <- tm_map(corpus, PlainTextDocument)
  corpus <- VCorpus(VectorSource(corpus))
  #BigramTokenizer <- function(x) NGramTokenizer(x, Weka_control(min = 2, max =
2))

```



```

BigramTokenizer <- function(x) unlist(lapply(ngrams(words(x), 3), paste, collapse =
" "), use.names = FALSE)

dtm <- DocumentTermMatrix(corpus,
                           control = list(tokenize=BigramTokenizer,
                                           weighting = weightTf))

rowTotals <- apply(dtm , 1, sum)

dtm.new <- dtm[rowTotals> 0, ]

lda <- LDA(dtm.new,number_of_topics,method =
'VEM',control=NULL,model=NULL)

topics <- tidy(lda, matrix = "beta") # convert the LDA output to a tidy
# get the top ten terms for each topic

top_terms <- topics %>% # take the topics data frame and..
  group_by(topic) %>% # treat each topic as a different group
  top_n(10, beta) %>% # get the top 10 most informative words
  ungroup() %>% # ungroup
  arrange(topic, -beta) # arrange words in descending informativeness

# if the user asks for a plot (TRUE by default)
if(plot == T){
  # plot the top ten terms for each topic in order


  top_terms %>% # take the top terms
  mutate(term = reorder(term, beta)) %>% # sort terms by beta value
  ggplot(aes(term, beta, fill = factor(topic))) + # plot beta by theme
  geom_col(show.legend = FALSE) + # as a bar plot

```

```
facet_wrap(~ topic, scales = "free") + # which each topic in a separate plot
labs(x = NULL, y = "Beta") + # no x label, change y label
coord_flip() # turn bars sideways
}else{
  # if the user does not request a plot
  # return a list of sorted terms instead
  return(top_terms)
}
}
```

top\_terms\_by\_topic\_LDA(X,T,6)

## APPENDIX B – Survey Questionnaire



**IOWADOT**  
I-80 Planning Study

I-80 Corridor Study

INTRODUCTION AND DEMOGRAPHICS

**The Iowa Department of Transportation is studying the Interstate 80 corridor and we would like you to take a few minutes to complete this survey. Your thoughts and opinions are important in helping the Iowa DOT make informed decisions regarding planning for the future of I-80.**

**Thank you, in advance, for your time and input.**

**1. What is your gender?**

Female  Male

**2. What is your current age?**

25 or under  26-35  36-45  46-55  56-65  Over 65

**3. What is the ZIP code for where you currently live? (Enter your five-digit ZIP code.)**

ZIP: