LITERATURE REVIEW OF TRAFFIC SAFETY JOURNAL PAPERS USING TEXT MINING

EDWARD RYAN CLAY – CAL POLY POMONA

WHAT IS TEXT MINING?

- A method that analyzes texts such as books and academic articles
- Determines word frequencies, word rank, sentimental values of words
- Topic analysis based on word pairs and their rankings

WHY USE TEXT MINING IN TRAFFIC SAFETY

- Countless Applications in Social Sciences and similar fields
- Numerous academic articles regarding traffic safety
- A different way to organize pre-existing articles

DATA USED

- Abstracts from Transportation Research Record
- Articles between 1996 and 2018
  - Easy to access
  - Large dataset

HOW TEXT MINING WAS PERFORMED

- Text Mining was performed utilizing R programing
- “tidy” method was used
- A number of packages used
- Data was separated into groups of 150 terms each

DATA DESCRIPTION

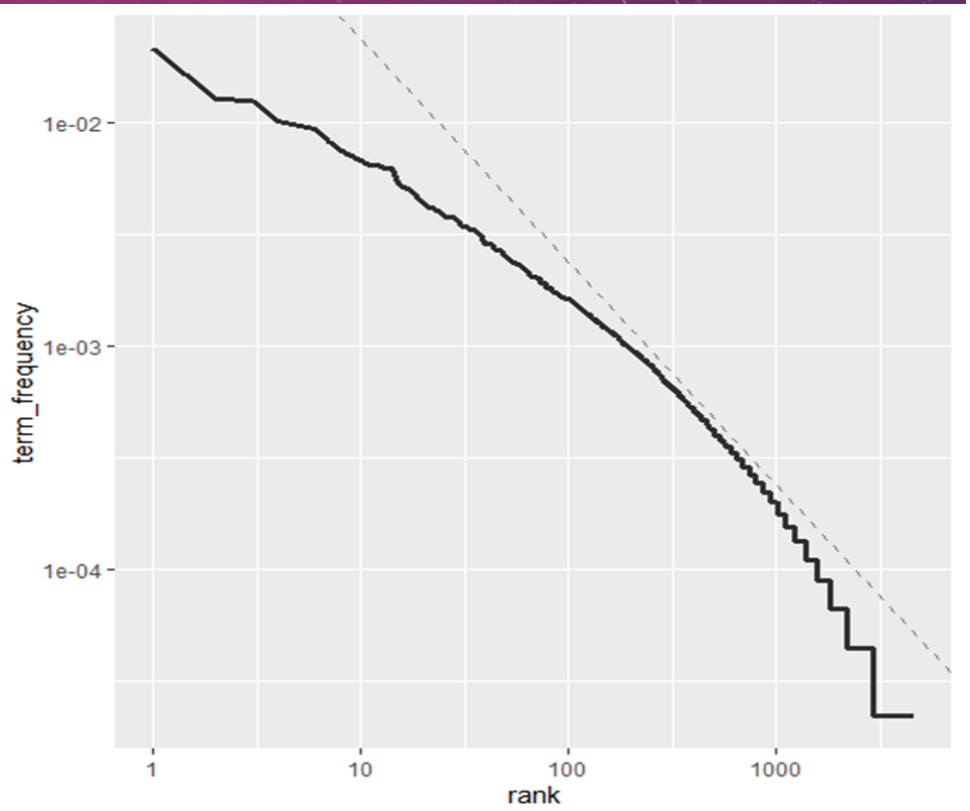
- About 450 abstracts from TRR alone
- No articles from 2001 and 2004

Year	Number of Abstracts	Words Counts		
		Minimum	Maximum	Mean
1996	9	82	248	157.33
1998	22	65	259	190.95
1999	13	73	441	191.31
2000	13	153	257	206.14
2002	18	126	269	199.22
2003	10	127	253	206.80
2005	25	61	298	198.92
2006	23	125	287	218.39
2007	30	117	269	207.69
2008	22	76	325	212.45
2009	28	115	281	211.39
2010	27	114	265	205.48
2011	30	139	271	216.60
2012	20	139	267	219.20
2013	21	166	260	219.81
2014	17	138	280	231.71
2015	19	186	290	241.26
2016	16	81	281	219.31
2017	57	118	269	217.49
2018	26	181	486	234.85

ZIPF'S LAW

- Term frequency isn't always directly correlated with term ranking
- Zipf's law shows that frequency is inversely related to ranking
- Zipf's law is true if slope is similar to -1

Note: the graph is on a logarithmic scale



SENTIMENTAL VALUE

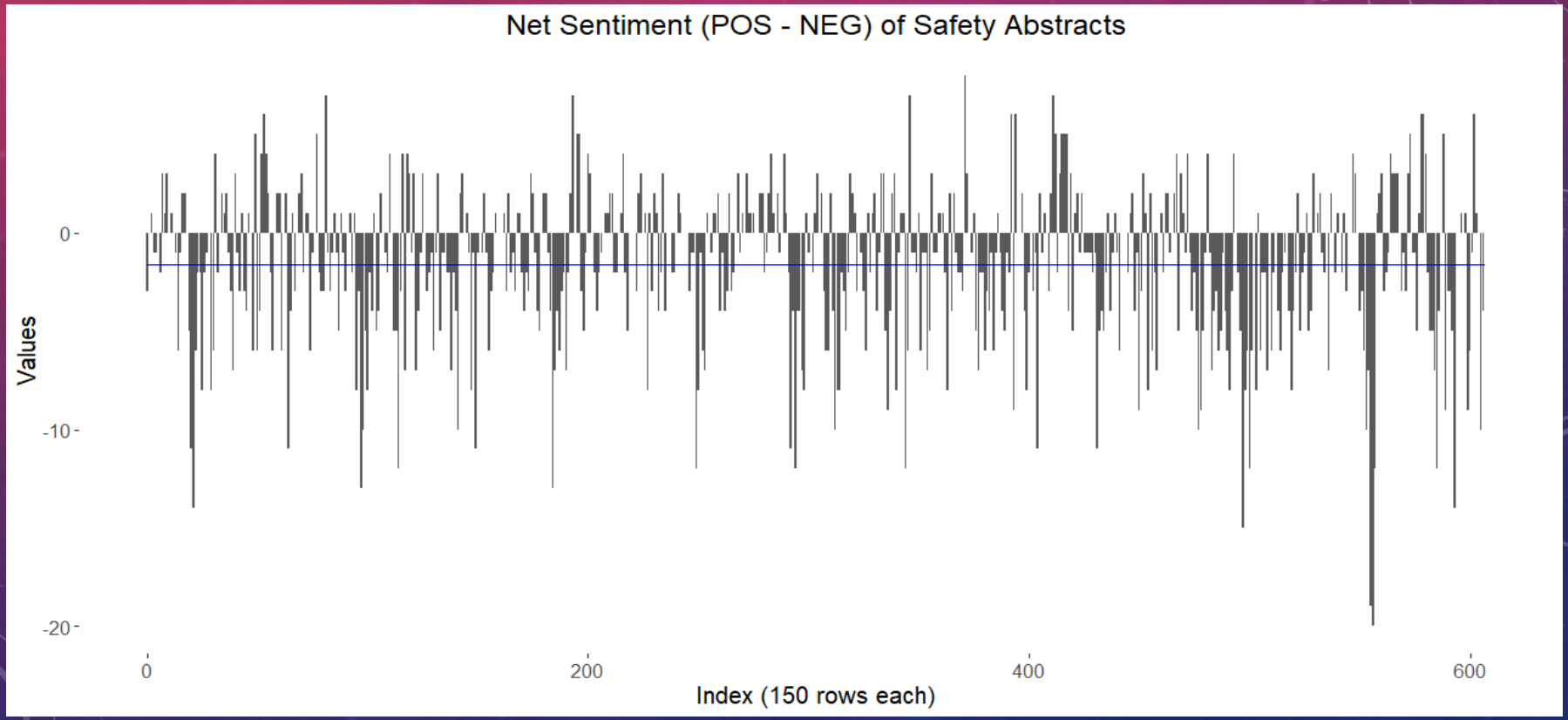
- “bing” method assigns a sentimental value to each word:
  - Negative sentimental words are -1
  - Positive sentimental words are +1

- “afinn” method assigns a sentimental value based on the sentimental intensity of each word
  - Negative sentimental words range from -5 to -1
  - Positive sentimental words range from +1 to +5

“BING” METHOD

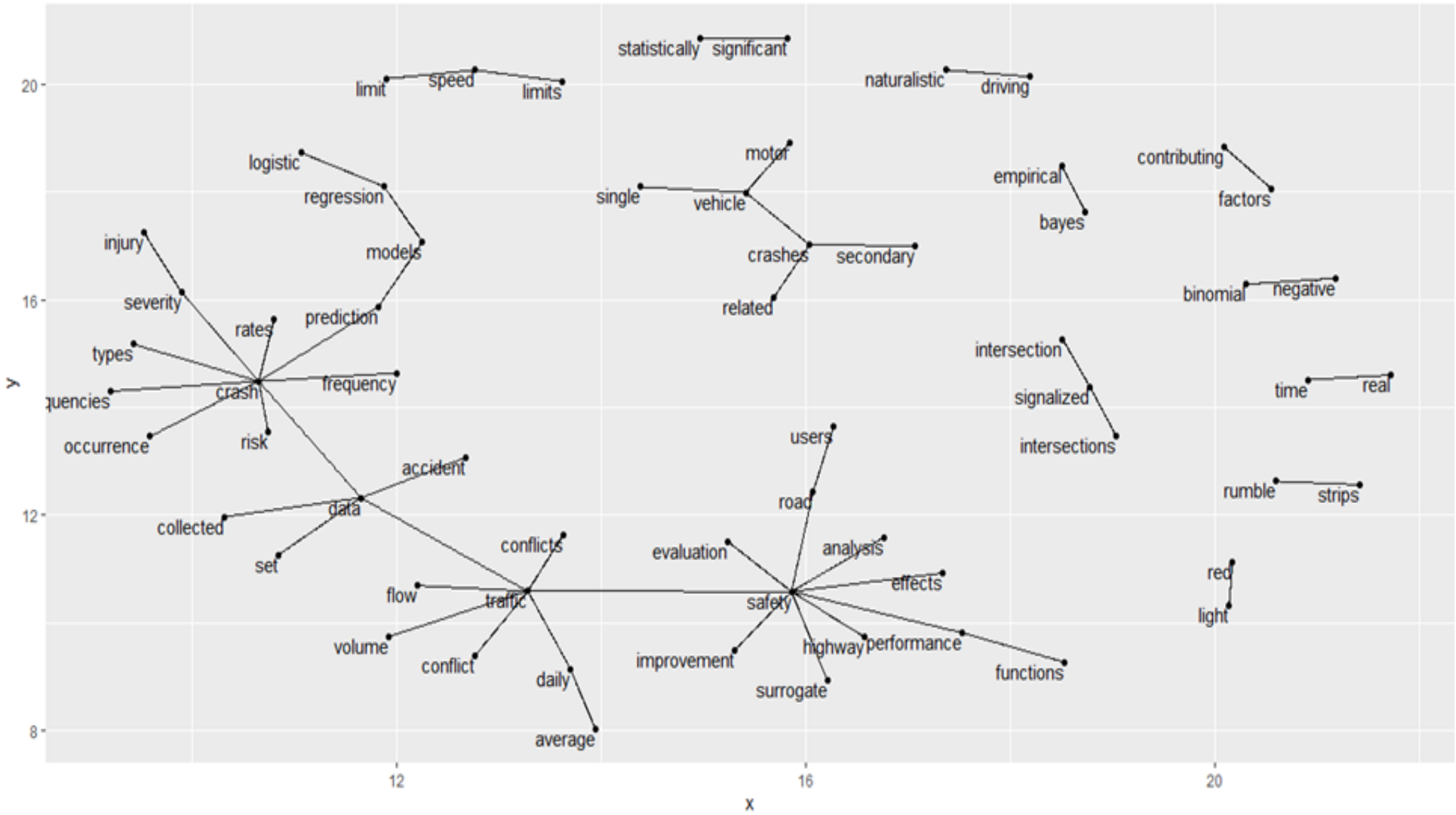
- Size of the word is related to frequency
- Color of word is related to sentimental value

“AFINN” METHOD



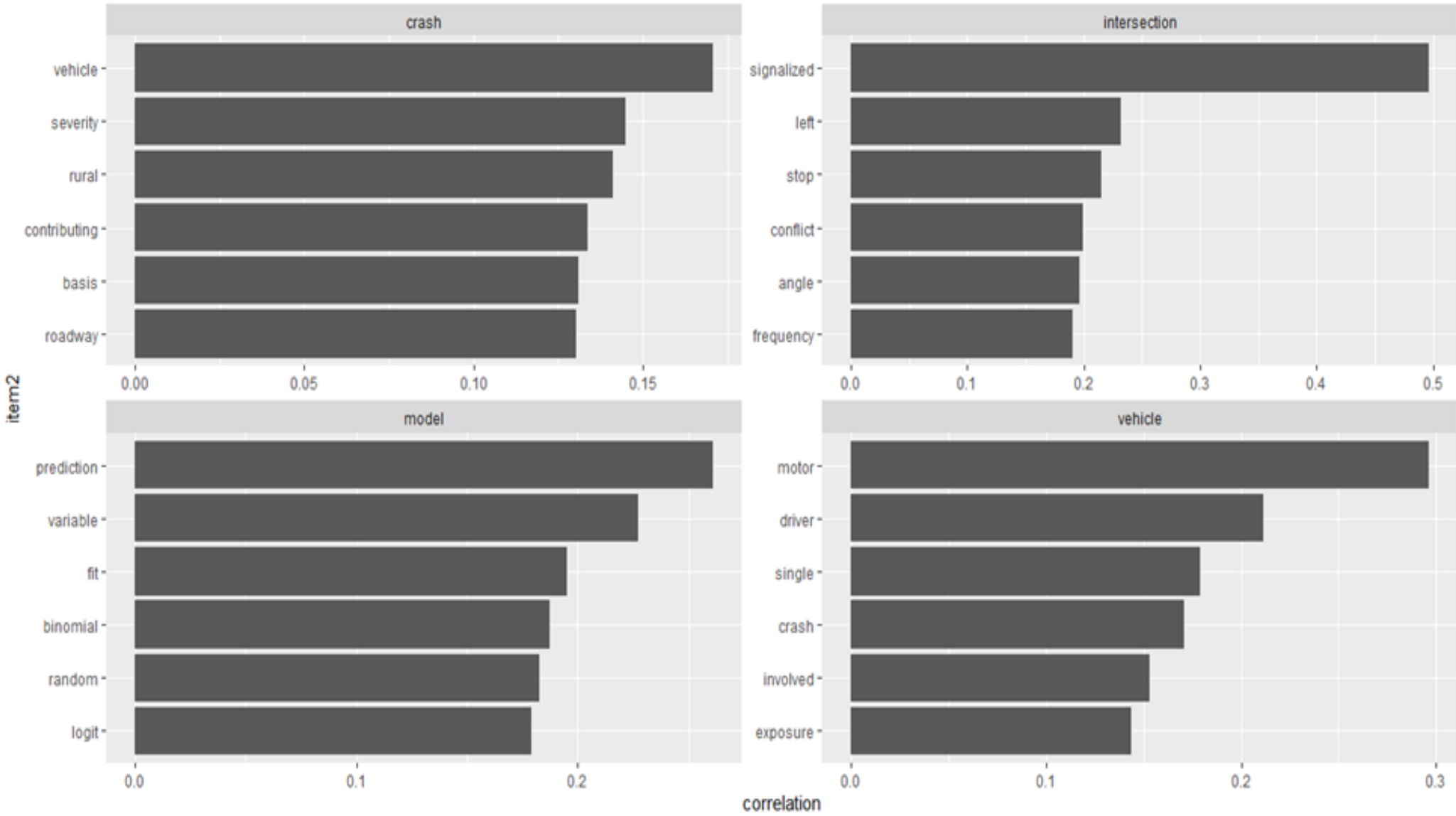
TOPIC ANALYSIS

- Certain words are more likely to appear in pairs
- Bigram analysis takes the frequency of these word pairs to give insight to the models used in traffic safety


statistically
significant
limit
speed
limits
naturalistic
driving
logistic
regression
motor
single
vehicle
empirical
bayes
contributing
factors
injury
models
crashes
secondary
severity
related
binomial
negative
prediction
rates
types
intersection
signalized
intersections
time
real
quencies
crash
frequency
occurrence
risk
users
accident
road
rumble
strips
data
collected
conflicts
evaluation
analysis
red
set
flow
traffic
safety
effects
light
volume
conflict
daily
improvement
highway
performance
functions
surrogate
average
20
16
12
8
y
12
16
20
x

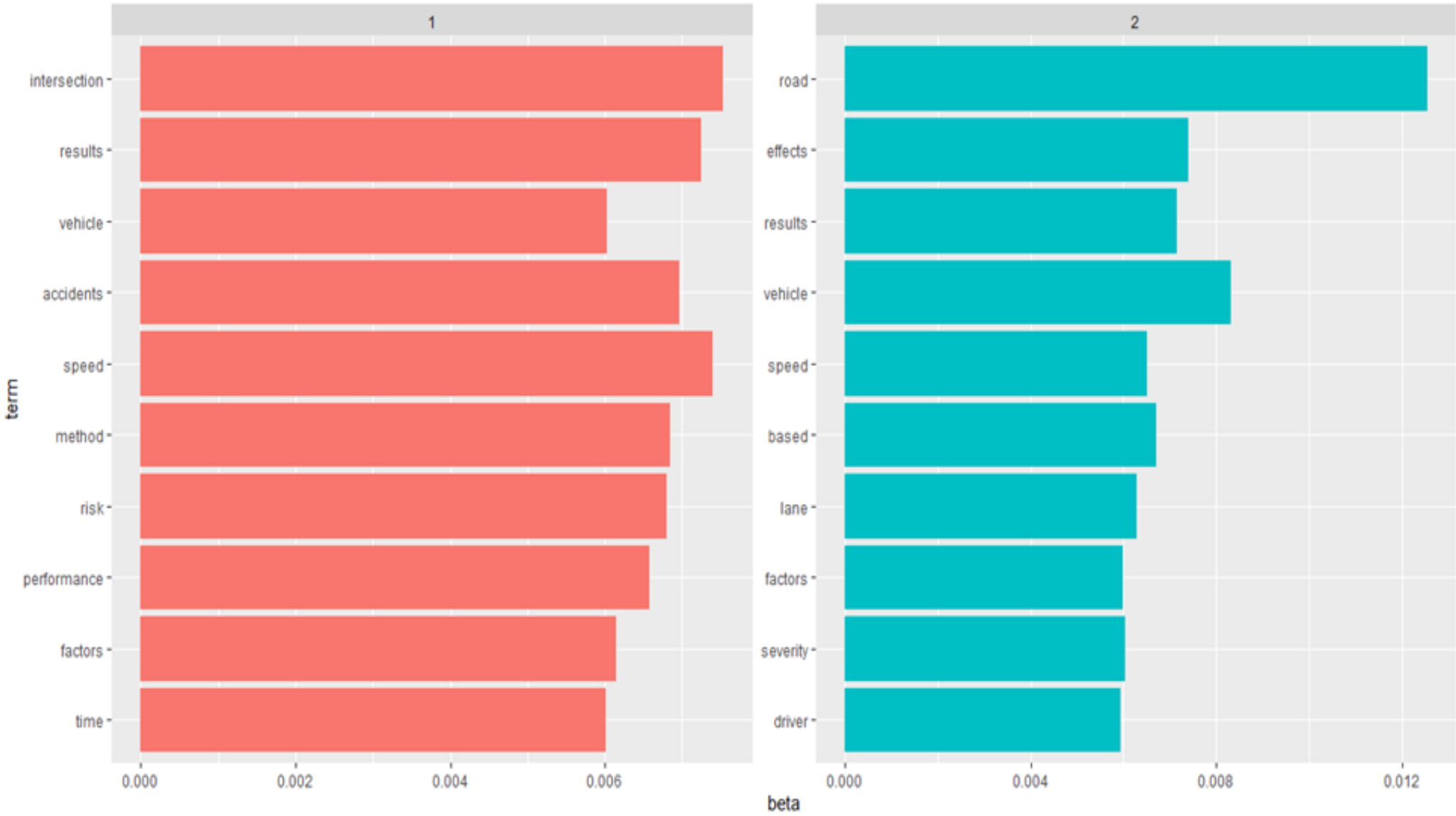
IMPORTANT WORD PAIRS

- Words such as “crash”, “intersection”, “model”, and “vehicle” are important in the field of traffic safety
- There is some overlap between the frequencies of each word pair


crash
vehicle
severity
rural
contributing
basis
roadway
0.00
0.05
0.10
0.15
intersection
signalized
left
stop
conflict
angle
frequency
0.0
0.1
0.2
0.3
0.4
0.5
model
prediction
variable
fit
binomial
random
logit
0.0
0.1
0.2
vehicle
motor
driver
single
crash
involved
exposure
0.0
0.1
0.2
0.3
item2
correlation

TWO MAJOR TOPICS

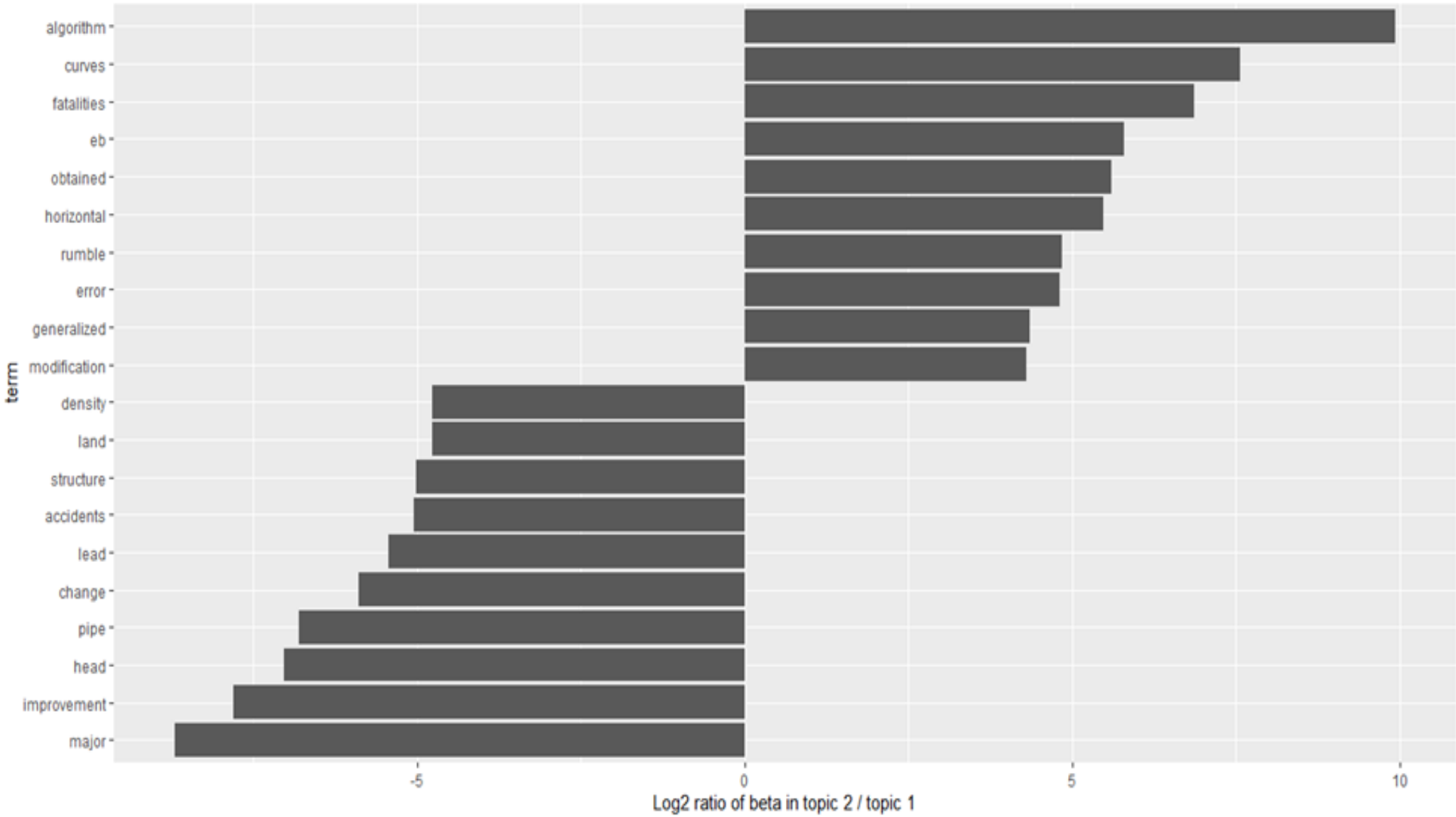
- Two major topics when analyzing data:
  - Intersection
  - Road
- Usually the two different places to analyze accident data
- The most common words within each topic on next slide


1
2
intersection
results
vehicle
accidents
speed
method
risk
performance
factors
time
road
effects
results
vehicle
speed
based
lane
factors
severity
driver
term
beta
0.000
0.002
0.004
0.006
0.000
0.004
0.008
0.012

LATENT DIRICHLET ALLOCATION (LDA)

- Another method of determining the major topics and their associated words
- Cross-over words that are related to both topics become more apparent to only one topic
- Results are on next slide

Note: The scale of the results is log base 2 and is a ratio of beta in topic 2 / topic 1


algorithm
curves
fatalities
eb
obtained
horizontal
rumble
error
generalized
modification
density
land
structure
accidents
lead
change
pipe
head
improvement
major
term
-5
0
5
10
Log2 ratio of beta in topic 2 / topic 1

CONCLUSIONS

- The overall sentimental value of the entire abstract data is negative according to both “bing” and “affin” methods
- The analysis of the word frequency and rank is inversely related which is verified by zipf’s law
- Correlation analysis within bigrams reveal that several statistical modeling methods such as negative binomial or topics such as signalized intersection are common areas of interest
- LDA proves to be a proper method to group topics in traffic safety.

FUTURE RESEARCH

- Another method other than “bing” or “afinn” to determine the sentimental value of text
- Understand the most common topics found in the data and potentially direct research to inadequately studied topics
- Pursue the different LDA methods to yield a more reliable analysis of topic modeling
- Broaden the data used, not only analyzing abstracts of TRR


THANK YOU!