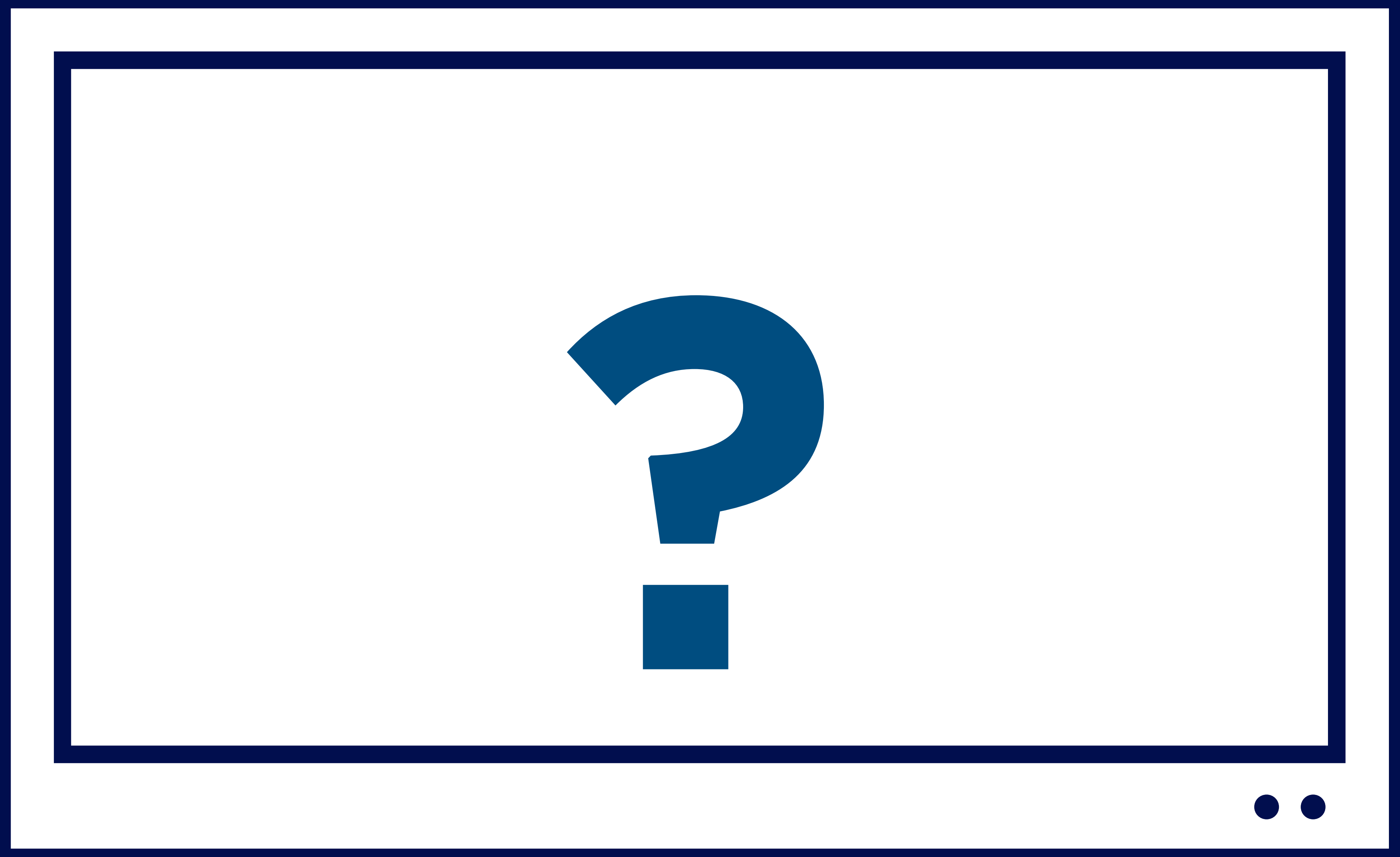
Text-Based Advertisement Feedback Topic Modeling

Springboard Data Science Bootcamp Capstone 2 Project
Presented by: Rebeca Mahr (Spring 2021)

Business Problem

Can a NLP machine learning model be developed to identify topics among text-based video ad feedback to inform message comprehension?



Standard practice is a manual review of feedback

- very time consuming
- subjective to reviewer

Image source: https://www.dreamstime.com/stock-illustration-businesswoman manager hold-long-list-scroll-tasks-questionnaire-woman-business-suit-flat character-vector image90834784

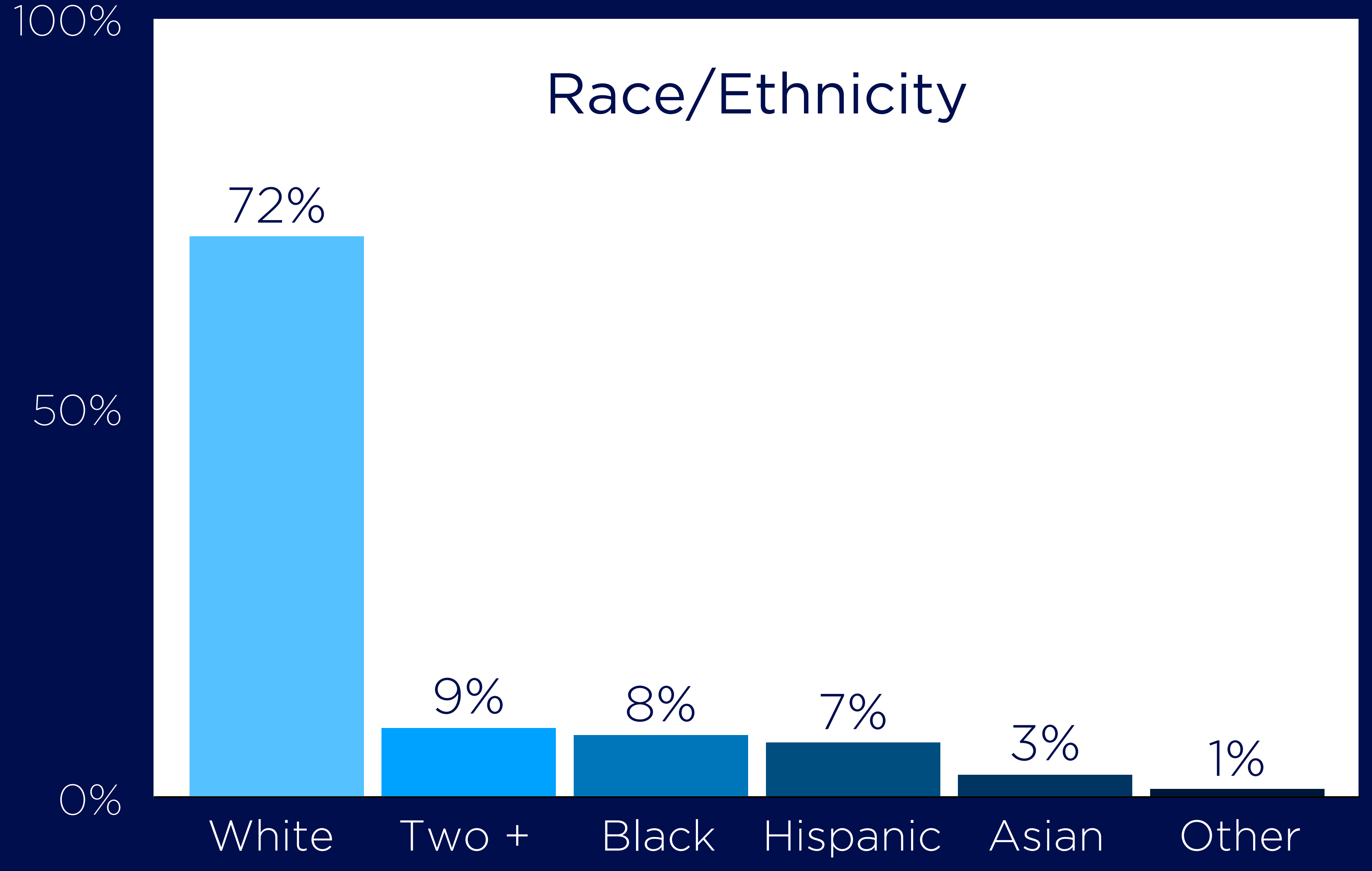
Data

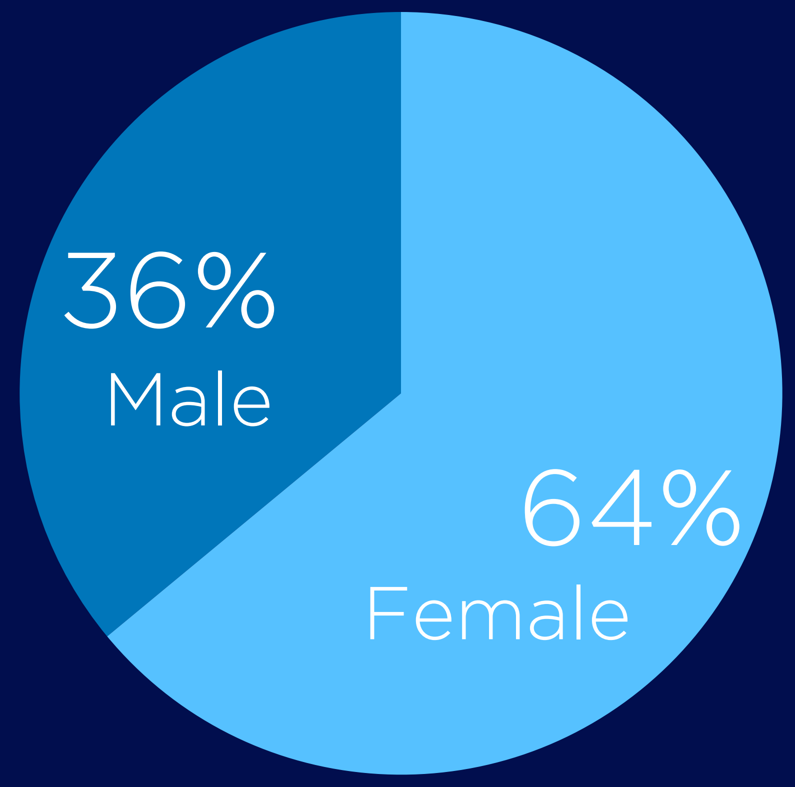
Ad Campaign	Nicotine-vape-prevention campaign
Target Audience	Teens within target US regions
Data Source	1,448 text-based survey responses to question “What do you think the main message of this ad is?” collected via Qualtrics online survey platform
Video Ads Tested	DD: message related to vape companies deceiving teens
	DF: message related to vapes making smokers vulnerable to viruses
	ST: message related to exposing the chemicals in vapes

n = 724

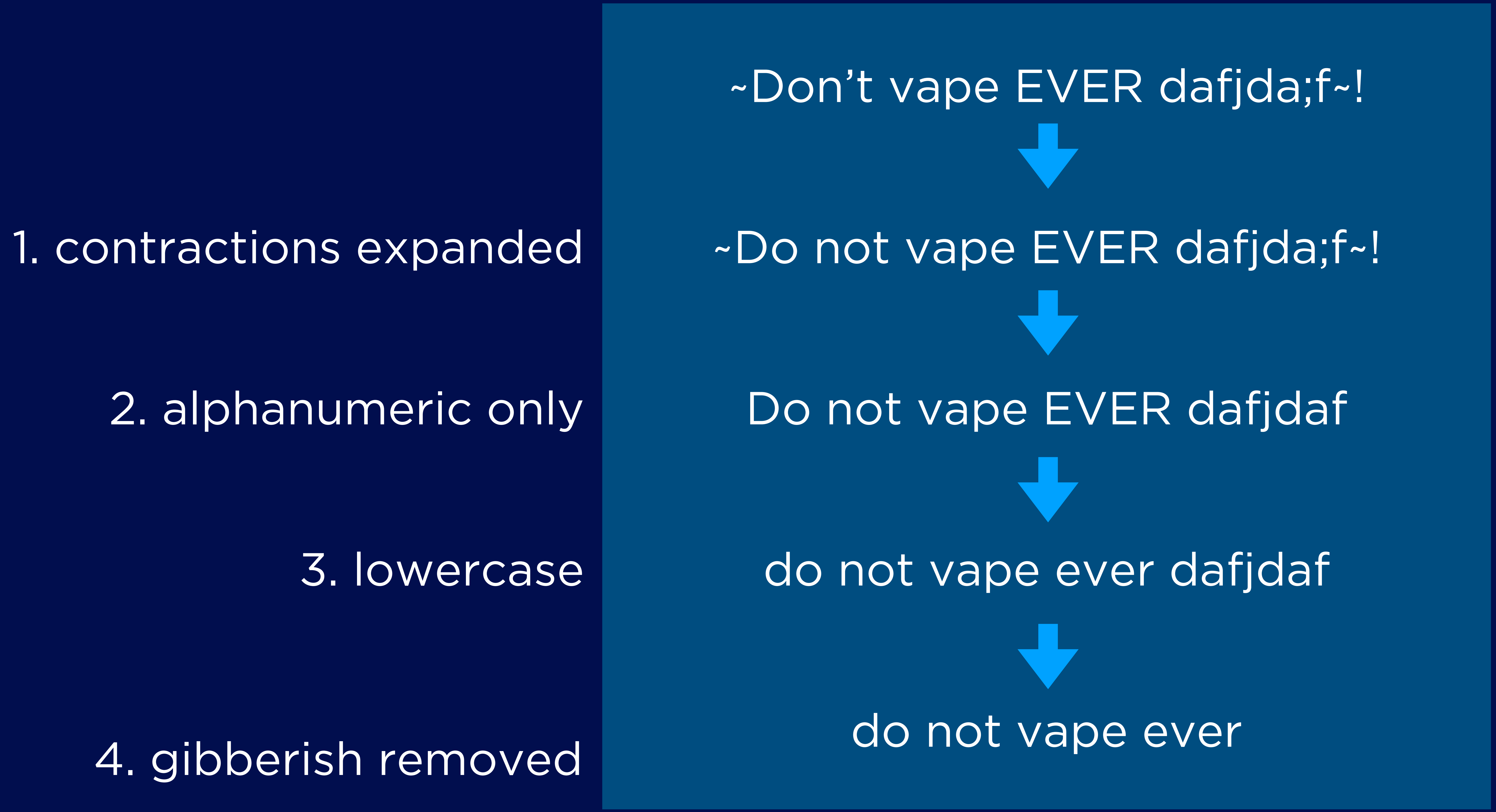
17

(avg. age)





Text Cleaning



~Don’t vape EVER dafjda;f~!

1. contractions expanded — ~Do not vape EVER dafjda;f~!

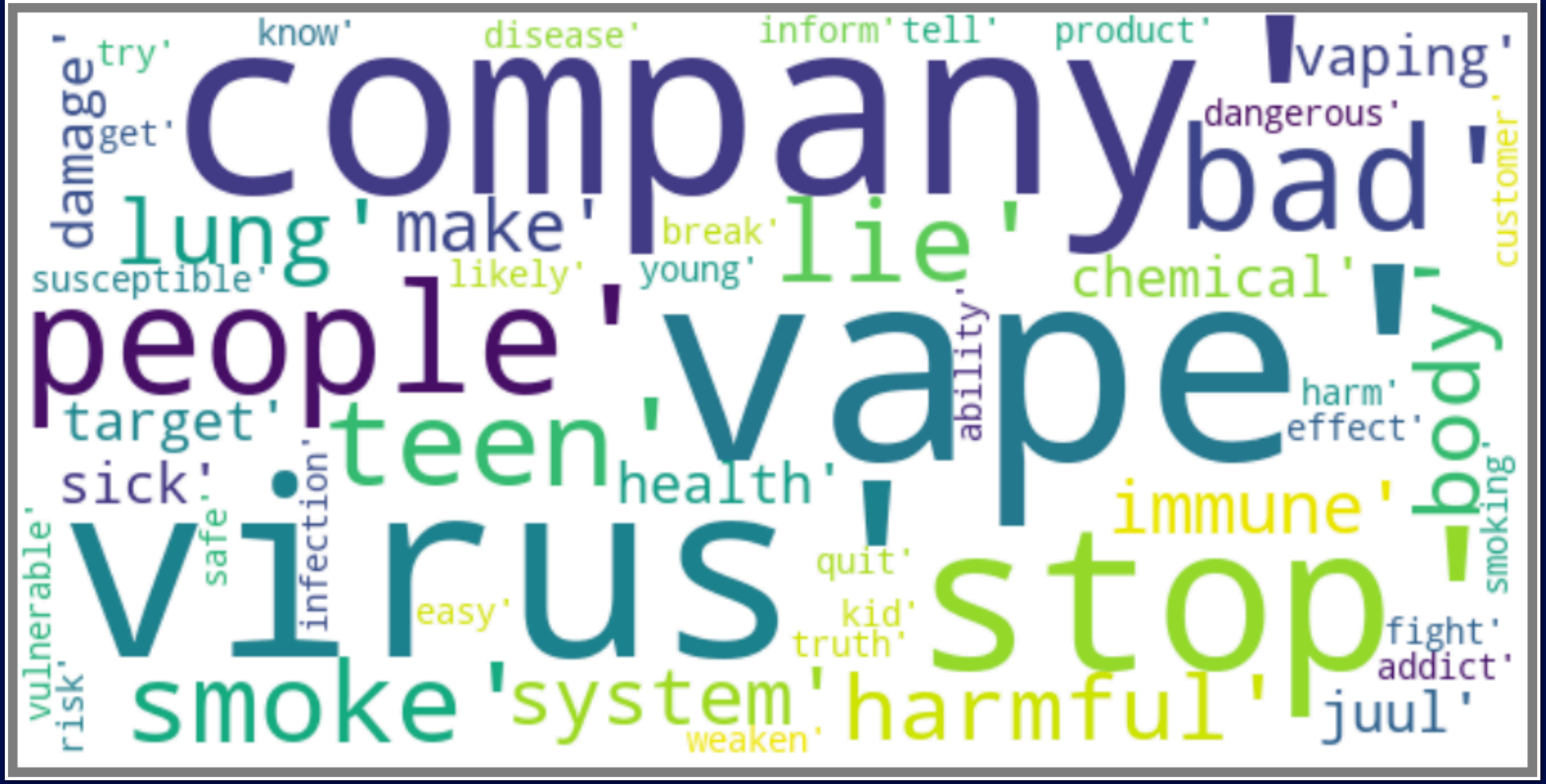
2. alphanumeric only — Do not vape EVER dafjdaf

3. lowercase — do not vape ever dafjdaf

4. gibberish removed — do not vape ever

Text EDA

Top 50 Words



Top 8 Words

vape 31%

company 5%

virus 4%

stop 4%

people 4%

bad 4%

teen 4%

smoke 4%

Preprocessing

Removal of stop words

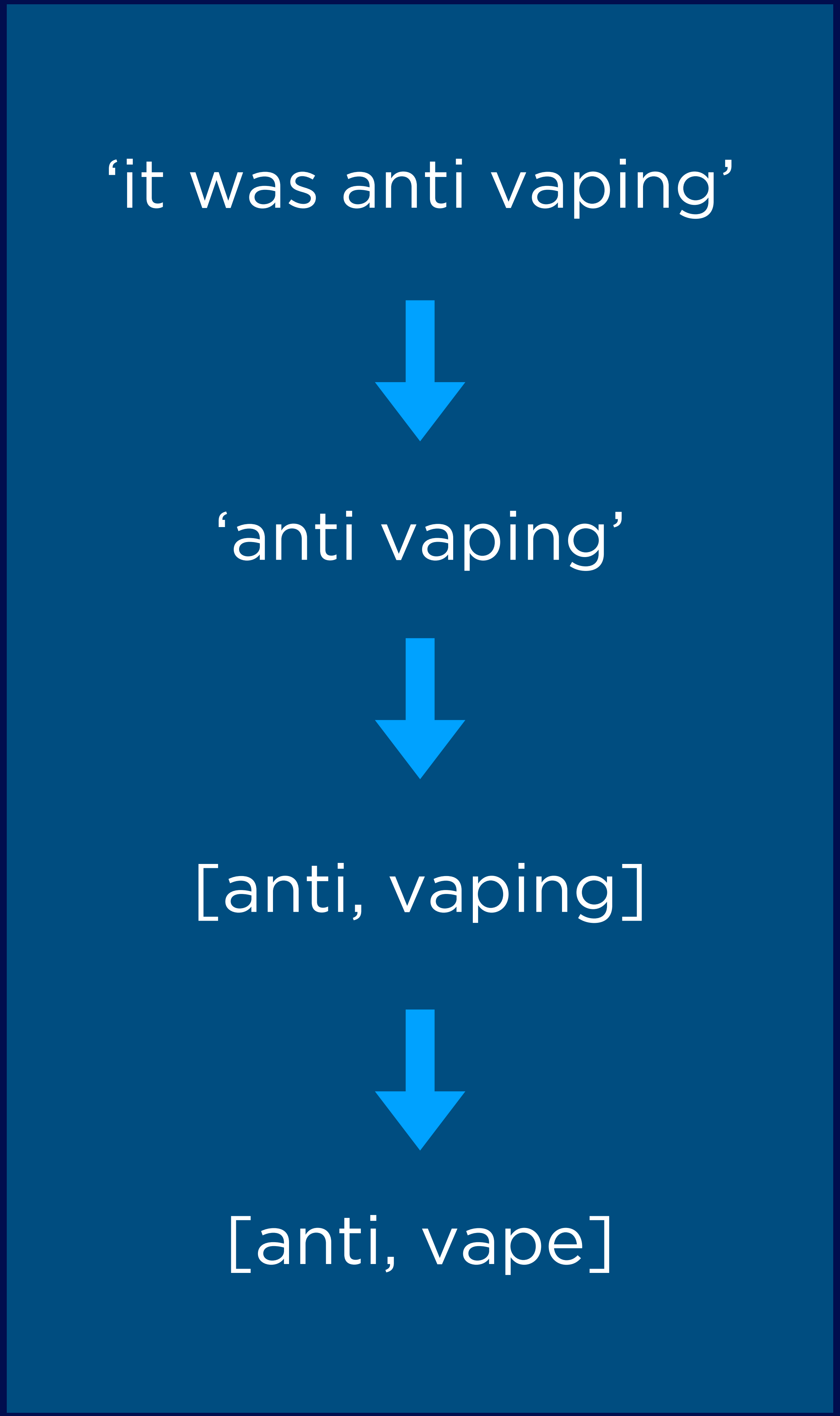
- examples: a, all, but, for, or, I, and...
- combination of Gensim, spaCy, and WordCloud stop words

Tokenization

- splitting text into meaningful tokens
- using spaCy Tokenizer

Lemmatization

- convert token words to root form
- using spaCy lemma_ method



‘it was anti vaping’

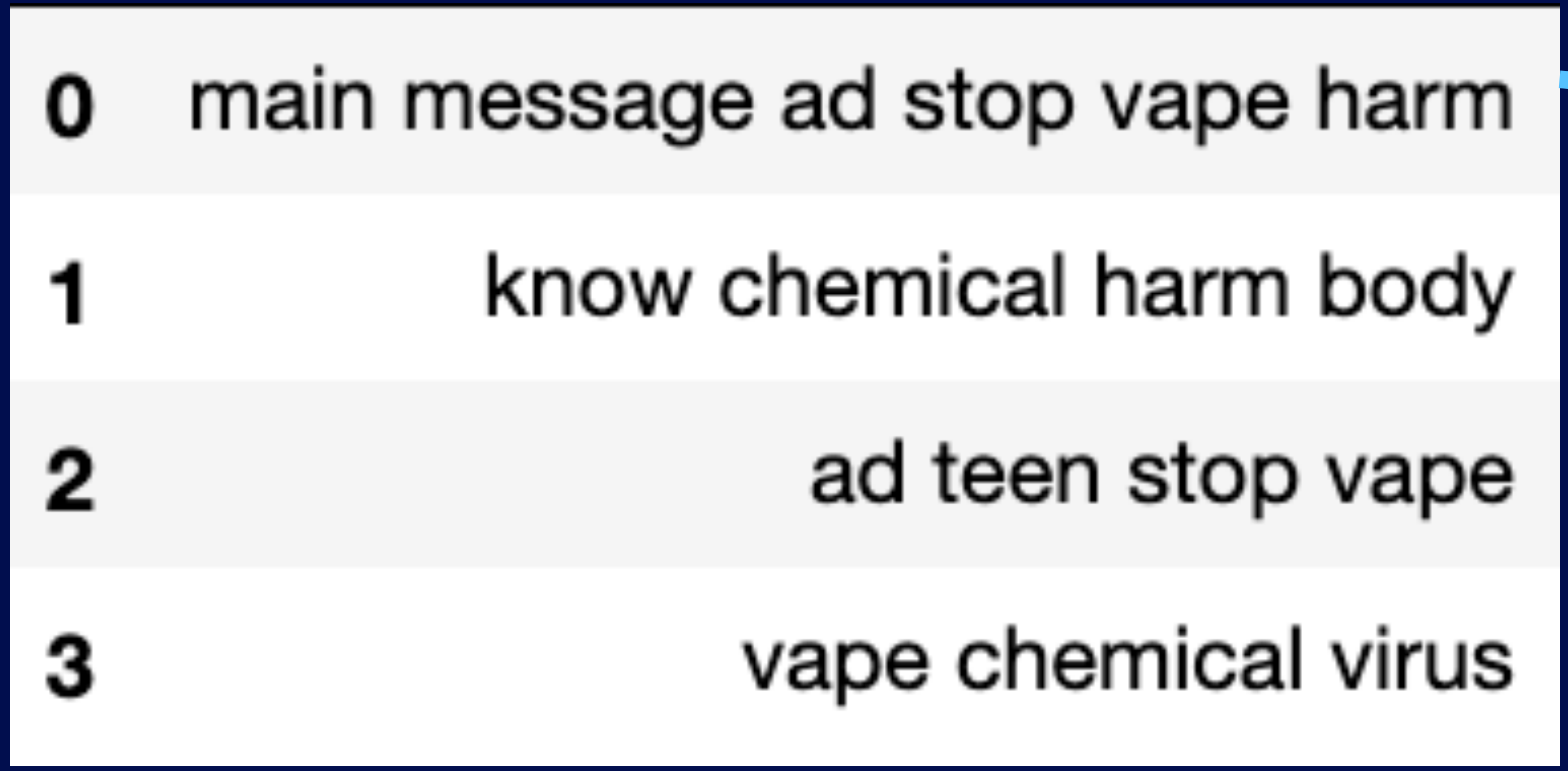
‘anti vaping’

[anti, vaping]

[anti, vape]

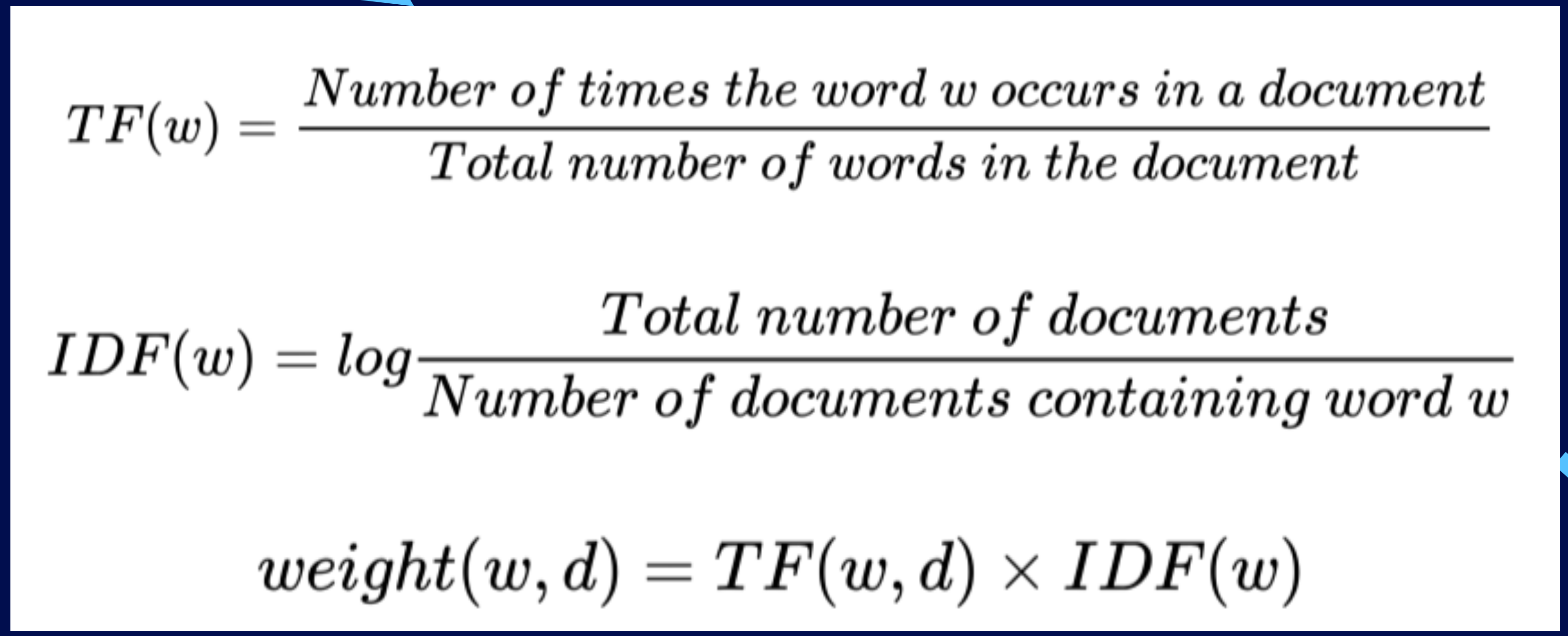
TF-IDF Vector

Processed Strings



0	main message ad stop vape harm
1	know chemical harm body
2	ad teen stop vape
3	vape chemical virus

weight terms



$$TF(w) = \frac{Number\ of\ times\ the\ word\ w\ occurs\ in\ a\ document}{Total\ number\ of\ words\ in\ the\ document}$$

$$IDF(w) = log\frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ word\ w}$$

$$weight(w, d) = TF(w, d) \times IDF(w)$$

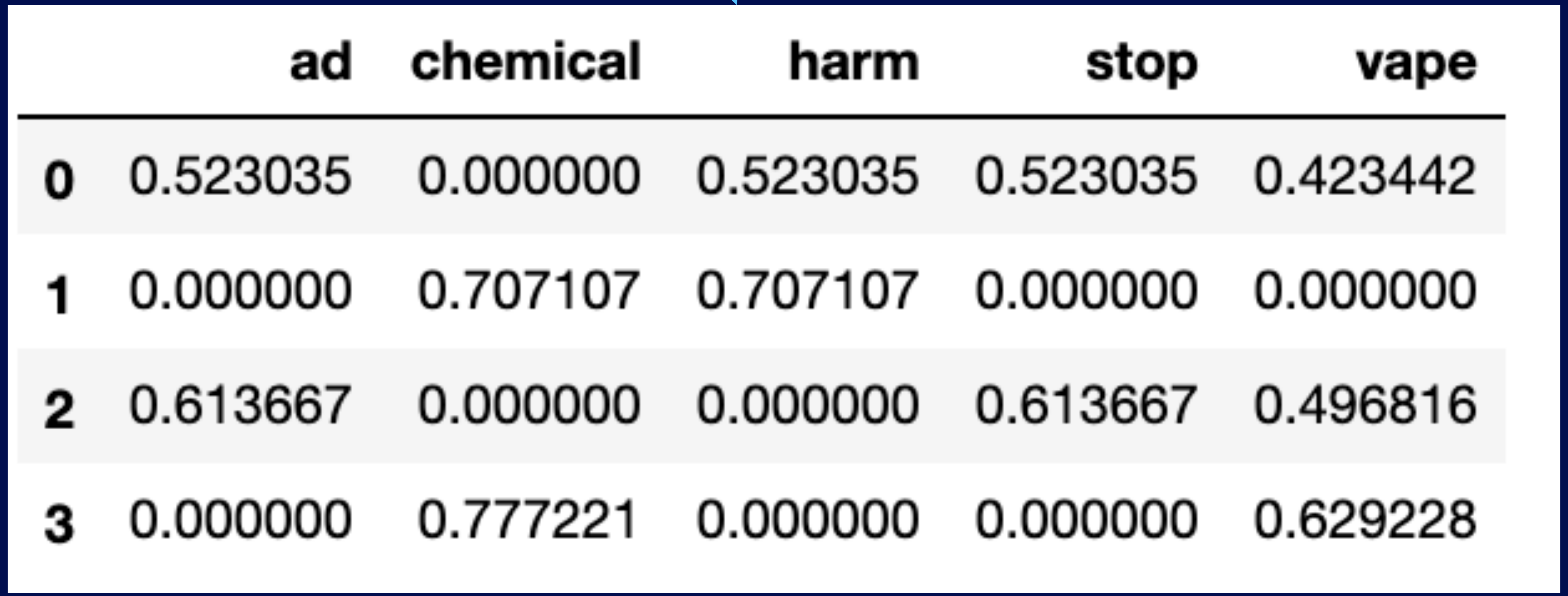
Formula source: Kedia, A., & Rasu, M. (2020). Understanding the Basics of NLP. In Hands on Python Natural Language Processing (p. 84). Birmingham - Mumbai: Packt Publishing Ltd.

max_df: .95

min_df: 2

use_id: True

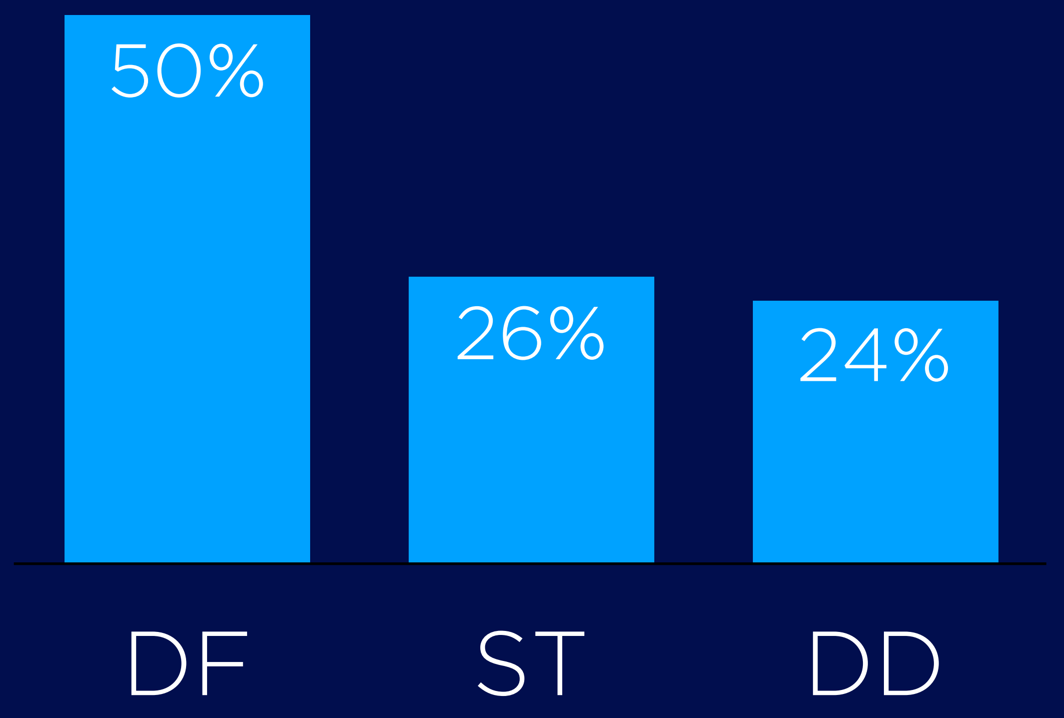
TF-IDF Vector



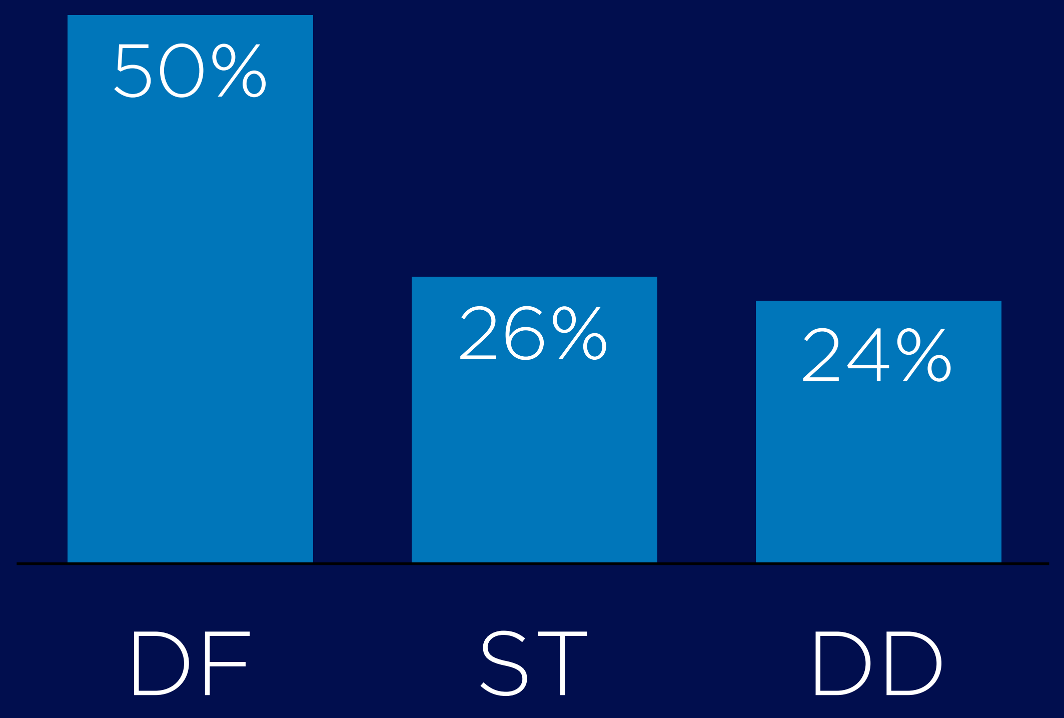
	ad	chemical	harm	stop	vape
0	0.523035	0.000000	0.523035	0.523035	0.423442
1	0.000000	0.707107	0.707107	0.000000	0.000000
2	0.613667	0.000000	0.000000	0.613667	0.496816
3	0.000000	0.777221	0.000000	0.000000	0.629228

Stratified Train / Test Split

70%



30%



Modeling

Initial LDA Topic Model

n_components: 6 topics (2 per ad)
max_iter: 250
learning method: online Bayes (for speed)

0	company teen lie vape target harmful juul product young try
1	smoke safe tell vape stop danger people inform inhale harmless
2	health risk disease vape high germ ingredient get spread increase
3	vape stop damage virus lung dangerous susceptible make vulnerable people
4	bad vape chemical body harm know good harmful people contain
5	immune weaken vape virus sick break fight easy likely body

Initial NMF Topic Model

n_components: 6 topics (2 per ad)
max_iter: 250

vape harmful dangerous chemical health immune know quit sick effect	0
bad chemical lung body health vape thing people know lot	1
stop people try body young kid inform encourage put help	2
company lie teen target juul product addict young kid people	3
smoke harmful dangerous health good cape vulnerable care inform kid	4
virus make lung susceptible damage immune body vulnerable weaken fight	5

Randomized Search Hyperparameter Tuning

Hyperparameter grid

- n_components: 3-12
- max_iter: 50-500 (increments of 50)

Other Parameters

- Iterations: 50
- cv: 5

Optimized LDA Topic Model

n_components: 3 topics
max_iter: 450

0	vape company teen harmful lie target dangerous sick juul harm
1	smoke stop virus vape people damage make lung susceptible vulnerable
2	bad vape immune weaken health know risk chemical quit tell

Optimized NMF Topic Model

n_components: 3 topics
max_iter: 450

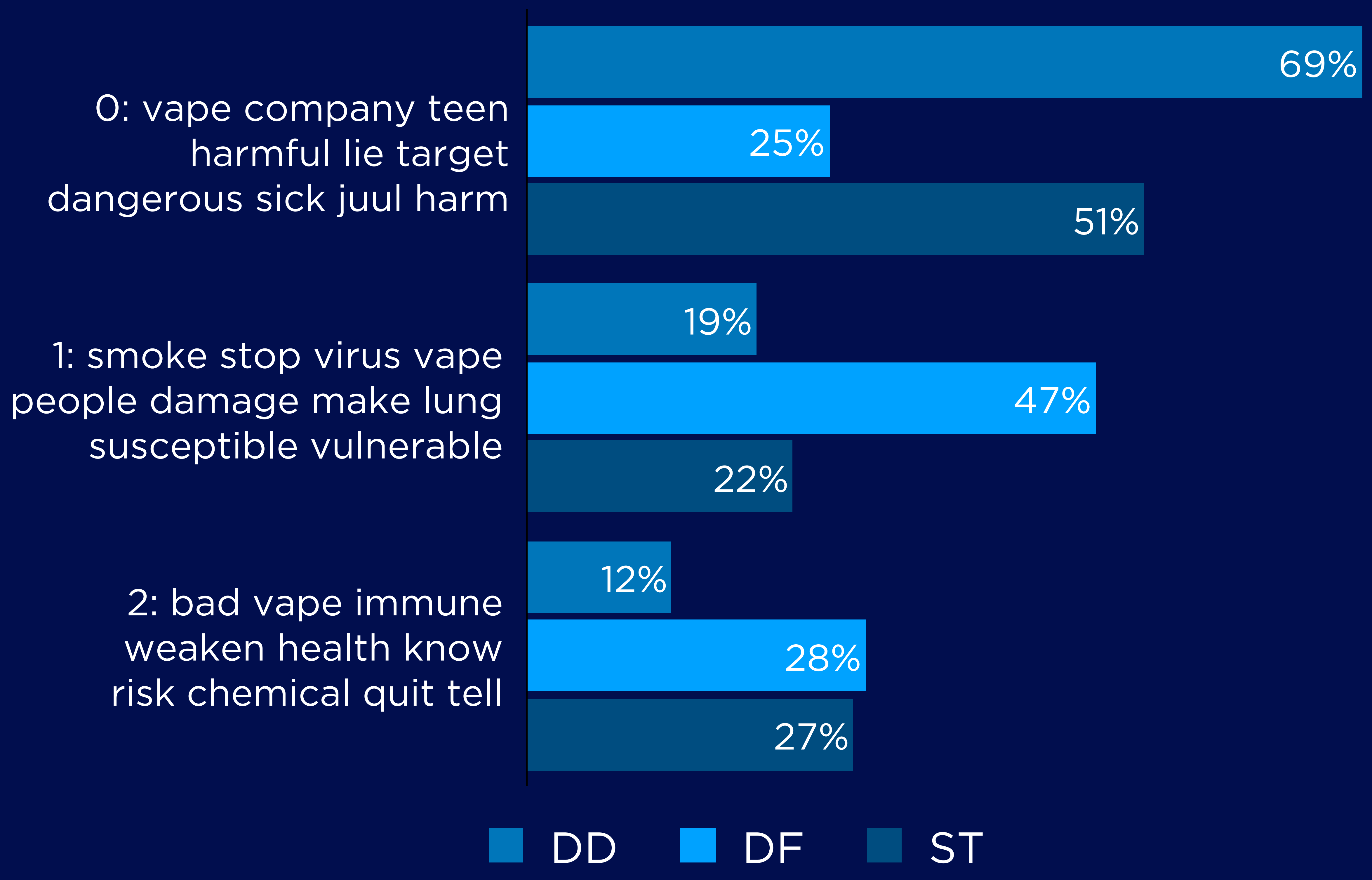
vape company harmful virus lie teen stop immune damage dangerous	0
bad vape chemical health smoke lung body thing know lot	1
smoke stop people health kid try good inform dangerous teen	2

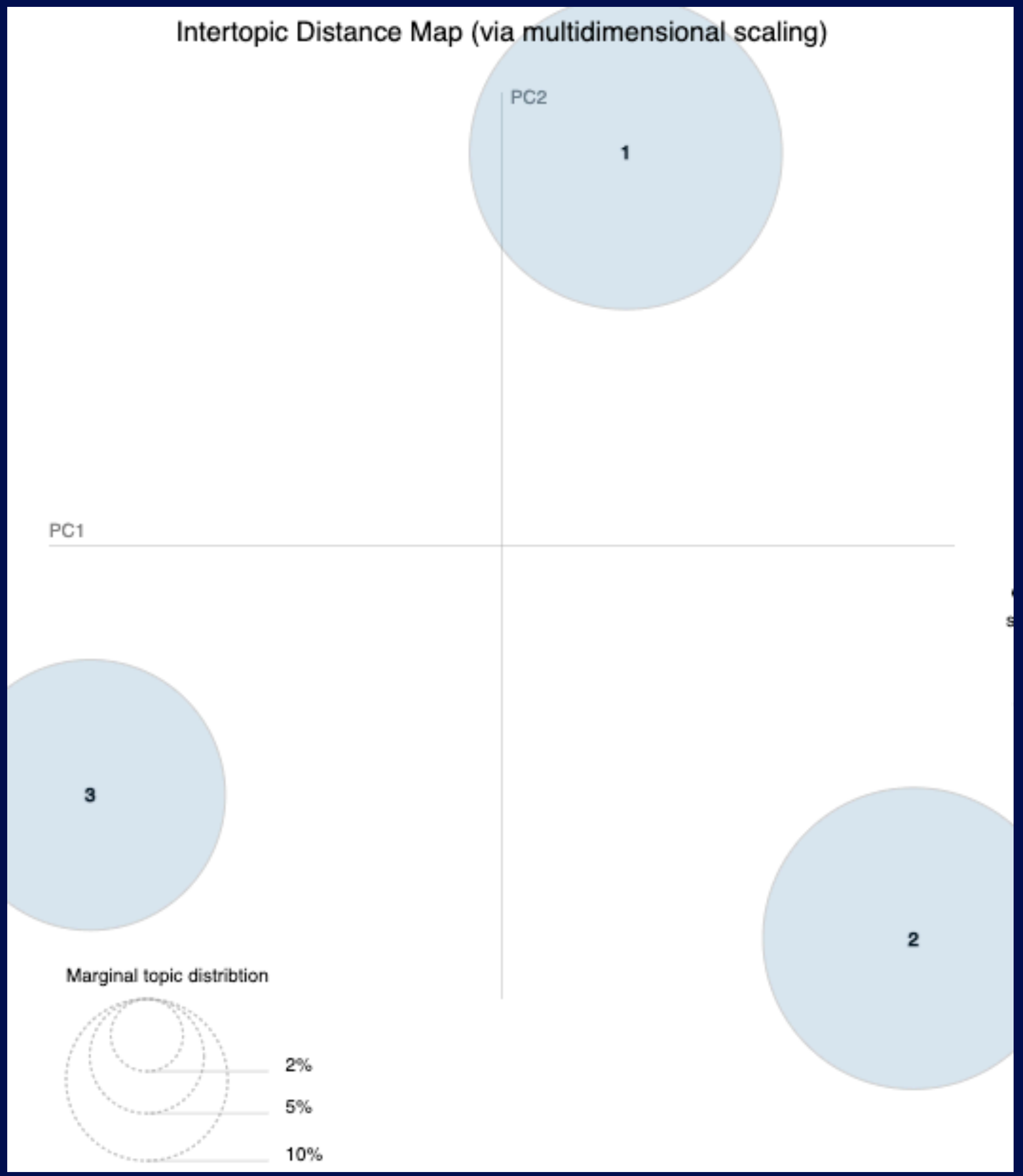
Selected Model

Optimized LDA Topic Model

- n_components: 3 topics
- max_iter: 450
- learning method: online Bayes (for speed)

- **51%** improvement in perplexity score





Conclusion

DD

message related to vape companies deceiving teens

69%

responses in topic “*vape company teen harmful lie target dangerous sick juul harm*”

✓

DF

message related to vapes making smokers vulnerable to viruses

75%

responses in topics “*smoke stop virus vape people damage make lung susceptible vulnerable*” and “*bad vape immune weaken health know risk chemical quit tell*”

✓

ST

ad message related to exposing the chemicals in vapes

51%

responses in topic “*vape company teen harmful lie target dangerous sick juul harm*”

✗

Limitations

- Sample selection bias
- Small sample size
- Skewed sample demographic distribution
- Subjectivity in model topic interpretation
- Limited options for NMF topic model evaluation
- Limited computational resources for hyperparameter tuning

Thank you!