

# Text Analytics

Keerthi Bojja | Mona Yao | Nimeelitha Akkiraju | Michael Wang | Pranidhi Prabhat

# Team



Keerthi Bojja



Mona Yao



Nimeelitha Akkiraju



Michael Wang



Pranidhi Prabhat

# Problem Statement

The data set comprises of Reddit's comments for different brands of cars. For a new user visiting the pages, the content is overwhelming at the first sight.

If he has a priority list for certain features in the automobile, he might not be able to reach the most valuable comments given the number of comments.

# Solution

To create a better user experience for these users, we planned to create filters of the most important topics under play in each of the brands comments, so that the user can filter out the comments related to these important topics.

# Content

- Approach
- Methodology
- Topic Modeling
  - LDA
  - Word2Vec
  - BERT

# Approach



---

Look for the keywords in the text corpus and use them to create topics

---

Determine the best approach to get the most relevant set of keywords

---

Explored different methods to finalize which methodology works the best and produces the most relevant filters

# Methodology

- Create a stop word corpus relevant to the dataset
- LDA – Latent Dirichlet Allocation
- Word2Vec – Create embeddings and use KMeans on the embeddings to create clusters
- BERT – Use BERT to create topic modeling using the UMAP and DBSCAN to establish the number of topics and then extract the top 5 ones

# LDA (Latent Dirichlet Allocation)

---



# LDA

- Documents -----> Topics -----> Words
- Matrix factorization technique

**Document – word matrix**

	W1	W2	W3	<u>Wn</u>
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
<u>Dn</u>	1	1	3	0

----->

	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
<u>Dn</u>	1	0	1	0

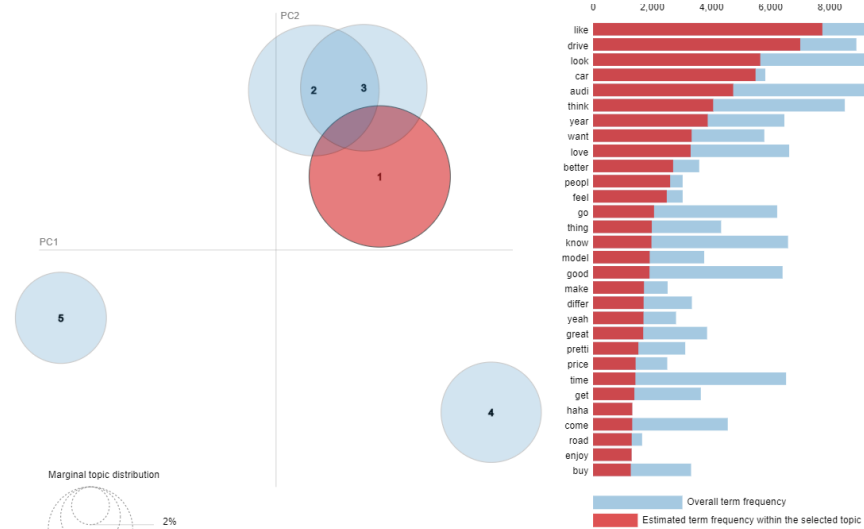
**Document - topic**

	W1	W2	W3	<u>Wm</u>
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

**Topic - word**

# Topics From LDA

Topic: 0  
Words:  $0.071 \cdot \text{"look"} + 0.035 \cdot \text{"delet"} + 0.033 \cdot \text{"wheel"} + 0.026 \cdot \text{"like"} + 0.026 \cdot \text{"black"}$   
Topic: 1  
Words:  $0.023 \cdot \text{"engin"} + 0.018 \cdot \text{"issu"} + 0.015 \cdot \text{"time"} + 0.014 \cdot \text{"tune"} + 0.012 \cdot \text{"chang"}$   
Topic: 2  
Words:  $0.057 \cdot \text{"audi"} + 0.056 \cdot \text{"http"} + 0.018 \cdot \text{"post"} + 0.016 \cdot \text{"model"} + 0.014 \cdot \text{"imgur"}$   
Topic: 3  
Words:  $0.028 \cdot \text{"like"} + 0.025 \cdot \text{"drive"} + 0.020 \cdot \text{"look"} + 0.020 \cdot \text{"car"} + 0.017 \cdot \text{"audi"} +$   
Topic: 4  
Words:  $0.027 \cdot \text{"thank"} + 0.018 \cdot \text{"audi"} + 0.014 \cdot \text{"work"} + 0.013 \cdot \text{"know"} + 0.012 \cdot \text{"tire"}$



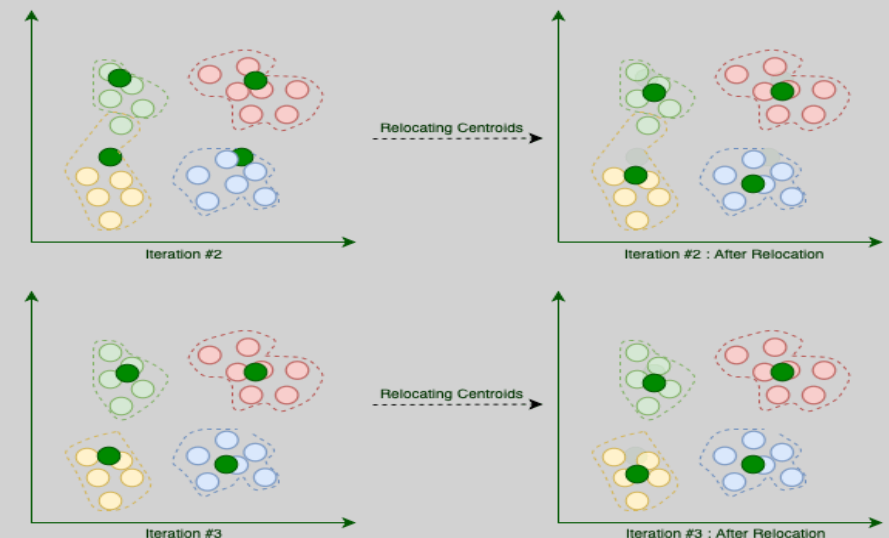
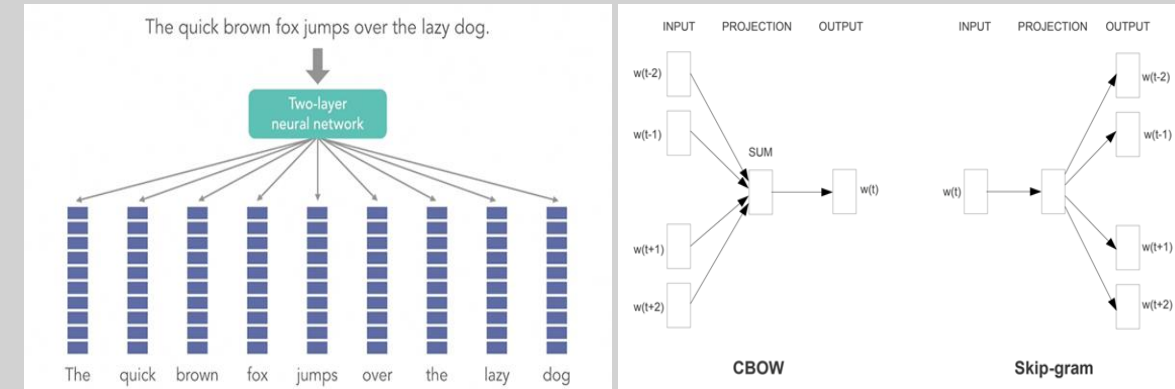
# Word2Vec



# Workflow

- Pre-processing
  - Stop words removal, stemming, lemmatization
- Word Embedding
  - Break down the cleaned text column into tokenized sentences
  - Use Word2Vec to extract different embeddings based on the context of the words
- Topic Modeling
  - Use K-Means to cluster the documents
  - Select optimal number of clusters using Elbow method & Silhouette Score

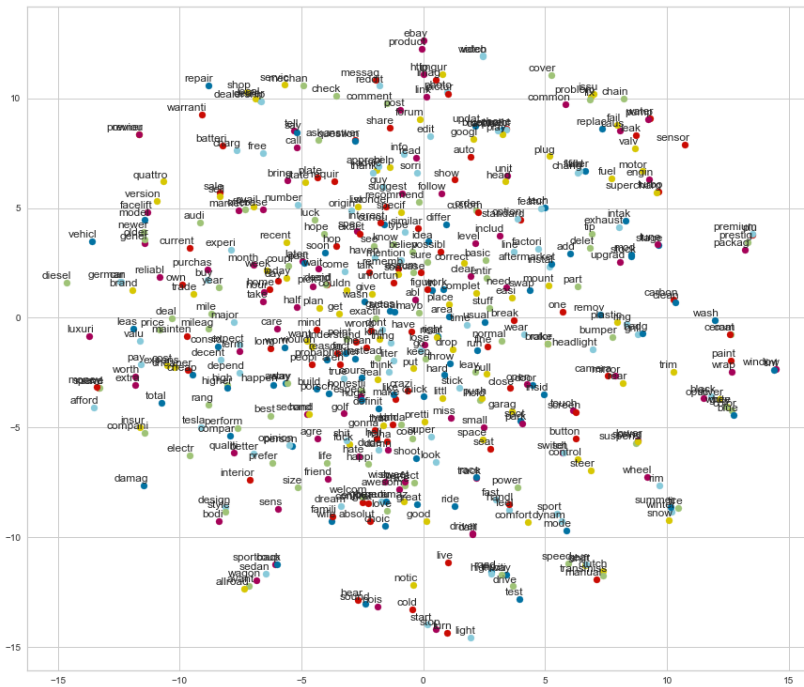
## Word2Vec



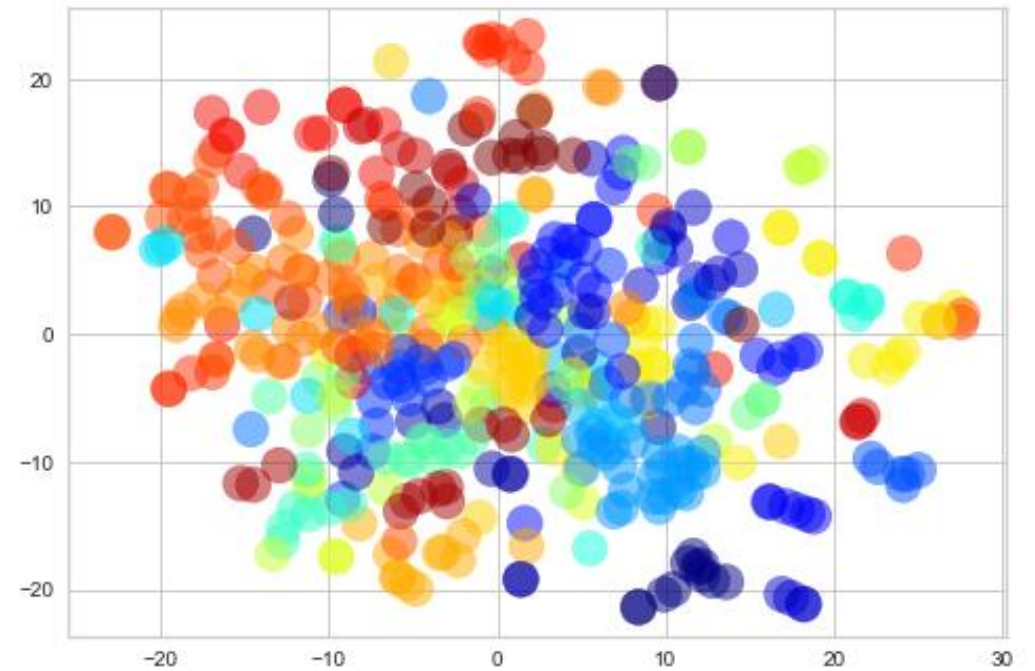
Termination: After Iteration #3 assignments won't change!

## K-Means

# Word2Vec-Embeddings



# K-Means Clusters



# Optimal Clusters

**Silhouette  
Score**

**Meaning**

1

Clusters are dense and nicely separated

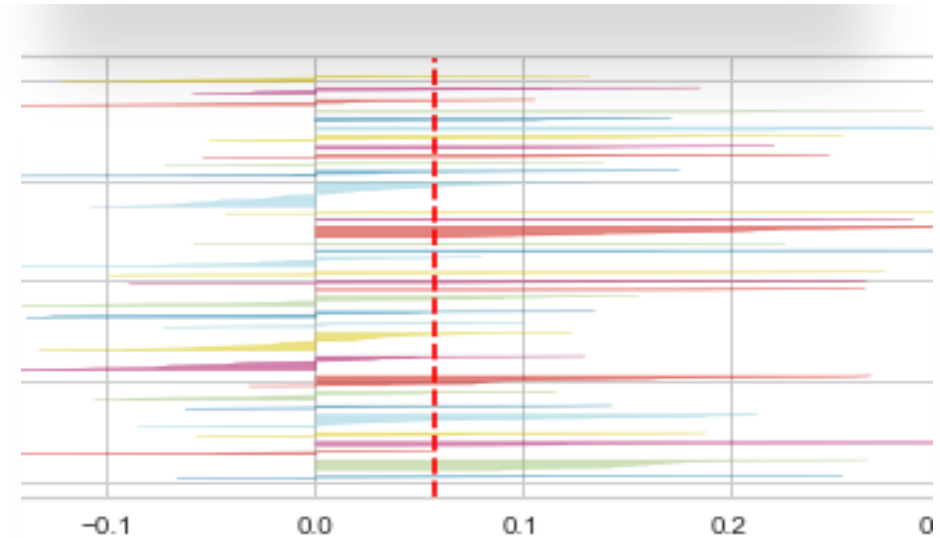
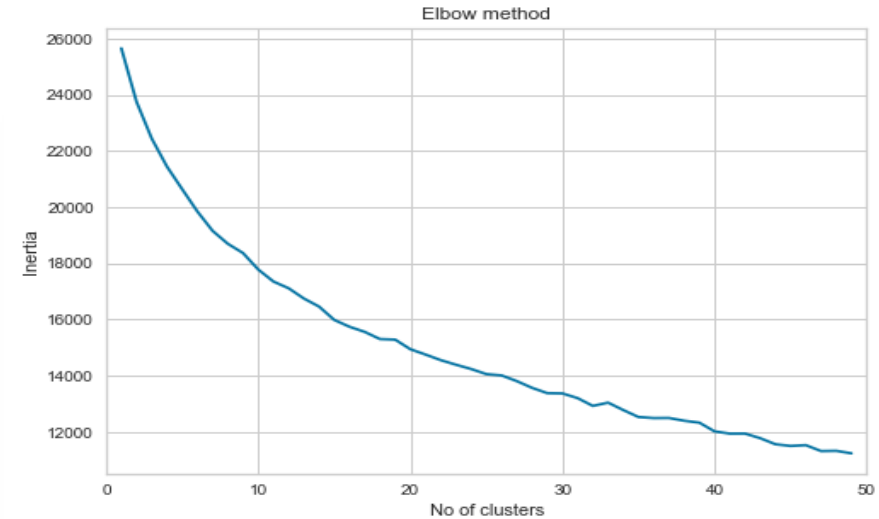
0

Clusters are overlapping

<0

Samples might have got assigned to the wrong clusters

## Elbow Method



Silhouette Score for 35 clusters: 0.068

# Word Cloud –Top 20 Words of each topic



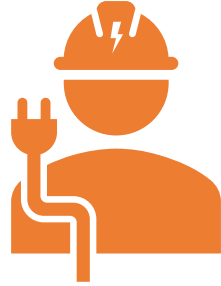
# BERT



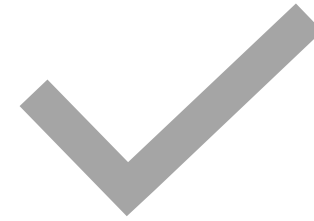


# Transformers

---



**Transformers provides general-purpose architectures (BERT, GPT-2, DistilBert etc.) for Natural Language Understanding and Natural Language Generation**



## **Sentence Transformers**

Turn text into BERT vectors

Cluster vectors into topics

Visualize topics

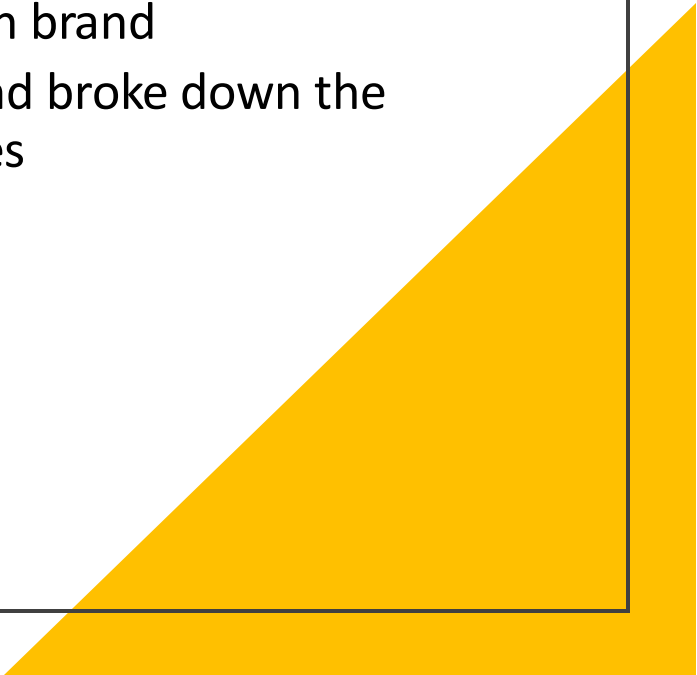
Check most frequently mentioned words within large topics

# What is BERT?

- BERT (Bidirectional Encoder Representations from Transformers)
- Difference between BERT and directional models
  - BERT reads entire sequence of word at once.
  - Allows to read the context of the word based on its surroundings
- The BERT model was pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables and headers).

# Data



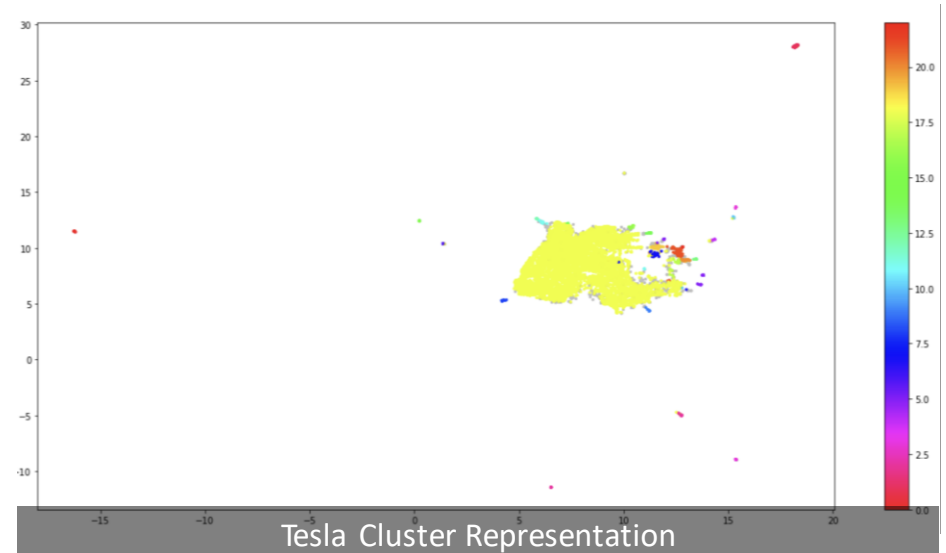
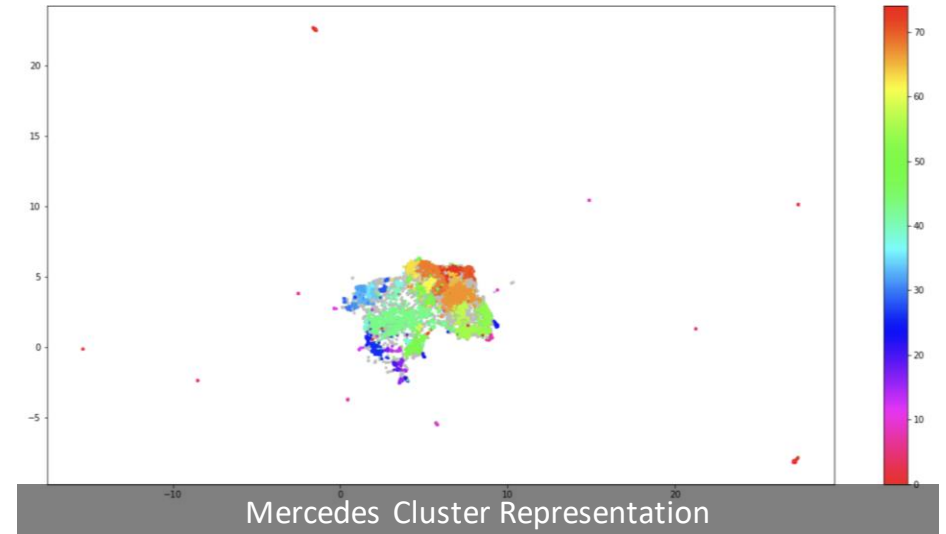
- Randomly selected 10k data points from the original dataset for each brand
  - Used column "body" and broke down the text body into sentences
- 

# Workflow

- Word Embedding
  - Break down the column "body" into sentences
  - Use BERT to extract different embeddings based on the context of the words
    - sentence-transformers library
- Clustering
  - Use UMAP to lower the dimensionality of the embeddings
  - Cluster the document using HDBSCAN
  - Reduce the dimensionality again to 2-D to plot the clusters
- Topic Creation
  - Use class-based TF-IDF to extract the single importance value for each word in a cluster
  - Use the top words in the cluster to create topic representations

# Clusters

- Some of the brands have relatively evenly spread topics
- The other brands have a very large topic



# Top Themes From UMAP Clustering

- BMW: dysfunction, Dyno, bhp (base horsepower)
- Audi: seat space, beams (lights), color (Nardo gray)
- Mercedes: warranty, AMG, aesthetics
- BoltEV: Chevrolet, “Http” (many links to images, videos)
- Leaf: Color, Social Media (YouTube),
- Tesla: Elon Musk, Social Media (YouTube), Defect

# Conclusion

LDA

Word2Vec

BERT

	C1	C2	C3	C4	C5
W1	look	engine	audi	like	thank
W2	delet	issu	http	drive	audi
W3	wheel	time	post	look	work
W4	like	tune	model	car	know
W5	black	chang	reddit	audi	tire

	C1	C2	C3	C4	C5
W1	entir	silver	unfortun	decent	fail
W2	basic	grey	sort	pretti	caus
W3	add	white	exactli	aren	fix
W4	assum	black	mention	near	problem
W5	curiou	color	assum	probabl	issu

	C1	C2	C3	C4	C5
W1	spot	based	space	seats	earlier
W2	shokan	lostredd itors	disasse mbled	leaving	closed
W3	cent	stage	tapes	visits	cassette
W4	nano	curb	wagons	watch	nardo
W5	thanks	3a	mpre	rover	thank

# Future Work

- Bert was run on CPU for this project and it is known that GPU runs it faster. If we were able to set up GPU properly, we can run the entire dataset instead of samples from each brand.
- Transformers provide many other interesting functionalities for text data, which we can explore