トヨタ自動車株式会社 御中

PQSコメント解析 Phase2 ―第二回―

2020年11月4日

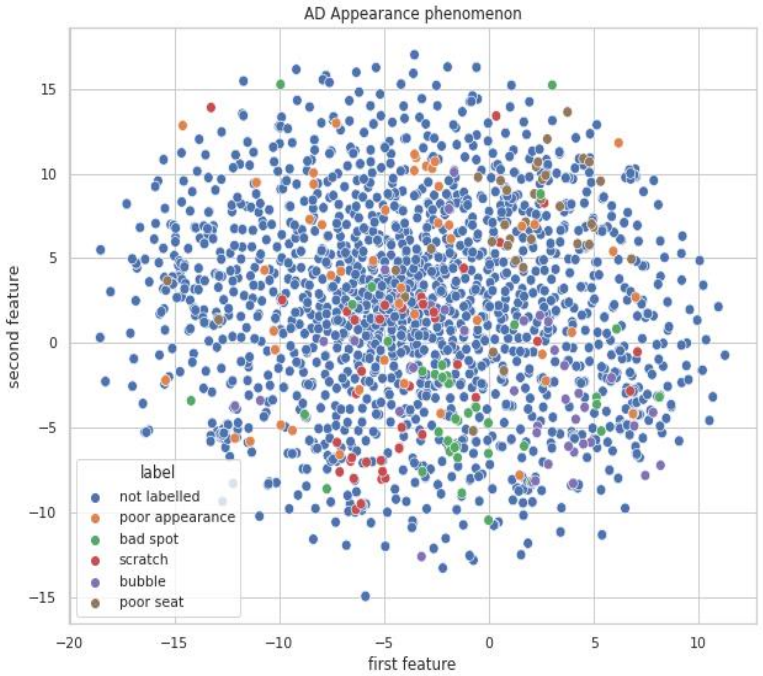
北大発
認定ベンチャー企業
Hokkaido University Venture

株式会社
調和技研

AD Appearance Annotation Visualization

Visualizing AD Appearance phenomenon using TF-IDF vectorization method

- Most of the area are covered by defined labels.
- But labels data points are overlapped each other.



Explaination

AD Appearance annotation coverage

- As we can see in the above graph, annotation coverage is well enough by the defined labels. Annotators estimated that 75% area are covered and we also observed that most of area are covered.
- We also noticed that, in the outer side of the above graphs there is only unlabeled data. So, Its true that nearly 25% are not covered as annotator estimated.
- To improve the coverage, we can increase the number of defined labels.

Explaination

AD Appearance annotation overlapping

- In the labeled data points, there is an overlapping problems for each of the labels.
- This is because of chosing generic labels for most of the items.
- Suppose, we select **poor appearance** but **poor appearance** belong to all most all the items in this phenomenon.
- Same goes for scratch, bad spot, and bubble.
- On the other hand **poor seat** is a specific label which represent only seat related comments. So, we can see, **poor seat** label's data point belongs to like a group in top-right corner of the graph.
- **Poor seat** label's comments does not overlap much like other four labels.

TF-IDF Vectorization method

Term Frequency Inverse Document Frequency(TF-IDF)

- We chose TF-IDF vectorization method
- TF-IDF formula
  - TF-IDF(t, D) = tf(t,d)*idf(t, D)
    - Tf(t,d) = $f_{t/d}$, Number of times appeared a term in particular document.
    - Idf(t, D) = log(N(total number of documents)/{number of document that appeared the term})
- Below are two documents(comments) for TF-IDF calculation.
  - Doc1: Rough spot in paint on hood.
  - Doc2: Rough spot on the driver side running board.

Vectorization method

Aplying TF-IDF method in two comments from AD pheonmenon

Doc1: Rough spot in paint on hood.
Doc2: Rough spot on the driver side running board.

Terms	TF		IDF	TF-IDF	
	Doc1	Doc2		Doc1	Doc2
rough	1	1	Log(2/2) = 0.0	0	0
spot	1	1	Log(2/2) = 0.0	0	0
in	1	0	Log(2/1) = 0.30	0.30	0
paint	1	0	Log(2/1) = 0.30	0.30	0
on	1	1	Log(2/2) = 0.0	0	0
hood	1	0	Log(2/1) = 0.30	0.30	0
the	0	1	Log(2/1) = 0.30	0	0.30
driver	0	1	Log(2/1) = 0.30	0	0.30
side	0	1	Log(2/1) = 0.30	0	0.30
running	0	1	Log(2/1) = 0.30	0	0.30
board	0	1	Log(2/1) = 0.30	0	0.30

Dimension reduction using PCA

Vectors for two documents

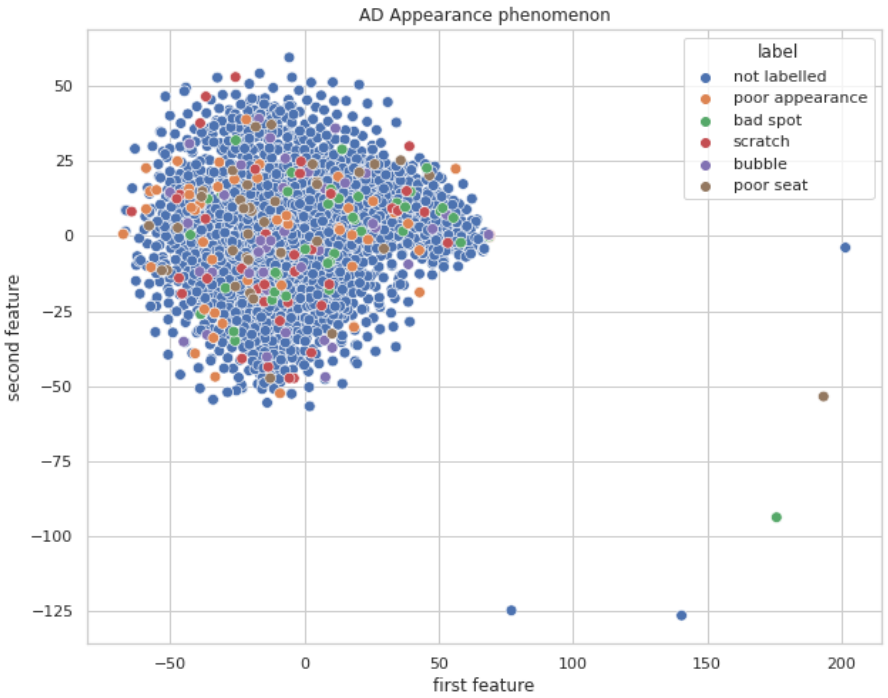
Doc1: [0.0, 0.0, 0.30, 0.30, 0, 0.30, 0.0, 0.0, 0.0, 0.0, 0.0]
Doc2: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.30, 0.30, 0.30, 0.30, 0.30]

- To visualize the document, we need to reduce the dimension of the vectors.
- For reducing dimension we used **T-SNE** which works on top of **Principal Component Analysis(PCA)**.
- PCA method filter the topmost important features from the vectors.
- After applying dimension reduction method in the above vectors, we got following vectors.
  - Doc1 : [3161.1775, 0.]
  - Doc2 : [-3161.1775, 0.]

AD Appearance Annotation Visualization

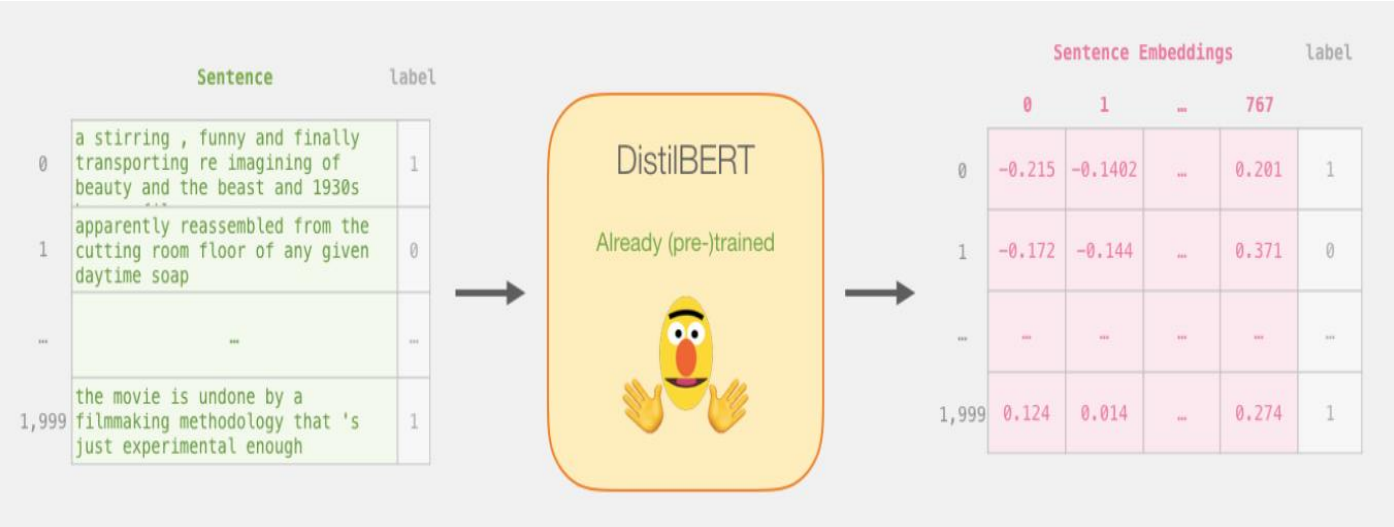
Visualizing AD Appearance phenomenon using BERT vectorization method

- Data points are more compressed compared to tf-idf method.
- Labels are covered the overall data points but label data points are overlapped each other. We explained the reason in previous graph.



BERT Vectorization Method

The BERT text vectorization architecture



- BERT has a complex architecture for vectorization.
- To convert each sentence to a tokenizer as list of words. And adding **CLS** token at the begining of the sentence. End of each sentence added **SEP** token**.**
- Each word considered as a token and assigned a token id.
- Converting the token list as a fixed length of token vector with adding zeros which called **padding**.

BERT Vectorization Method

The BERT text vectorization architecture

- Prepared another input vectors which is called M**asking**. Converting non-zero value to 1 so that model can focuse only only non-zero value. Thats why its called **Attention Mask**.
- Taking the tokenization vectors and attention mask as input to the BERT model and pass it to the intermediate layers. And in the **last hidden layer**, it contains 768 length vector. We extracted it and reduced the dimension using PCA for plotting.

BERT Vectorization Method

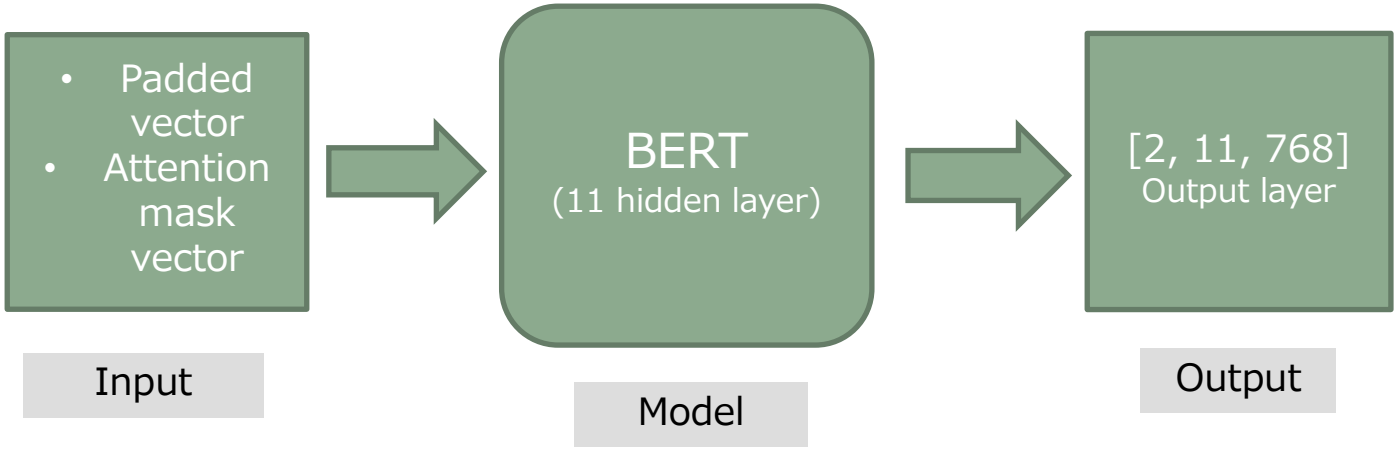
Applying BERT vectorization for two comments

Doc1: Rough spot in paint on hood.
Doc2: Rough spot on the driver side running board.

- After applying tokenizer for tokenization vector.
  - [101, 5931, 3962, 1999, 6773, 2006, 7415, 1012, 102]
  - [101, 5931, 3962, 2006, 1996, 4062, 2217, 2770, 2604, 1012, 102]
- After applied padding for converting padded vectors.
  - [101, 5931, 3962, 1999, 6773, 2006, 7415, 1012, 102, 0, 0]
  - [101, 5931, 3962, 2006, 1996, 4062, 2217, 2770, 2604, 1012, 102]
- To make mask vectors as attention mask.
  - [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0]
  - [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

BERT Vectorization Method

Applying BERT vectorization for two comments

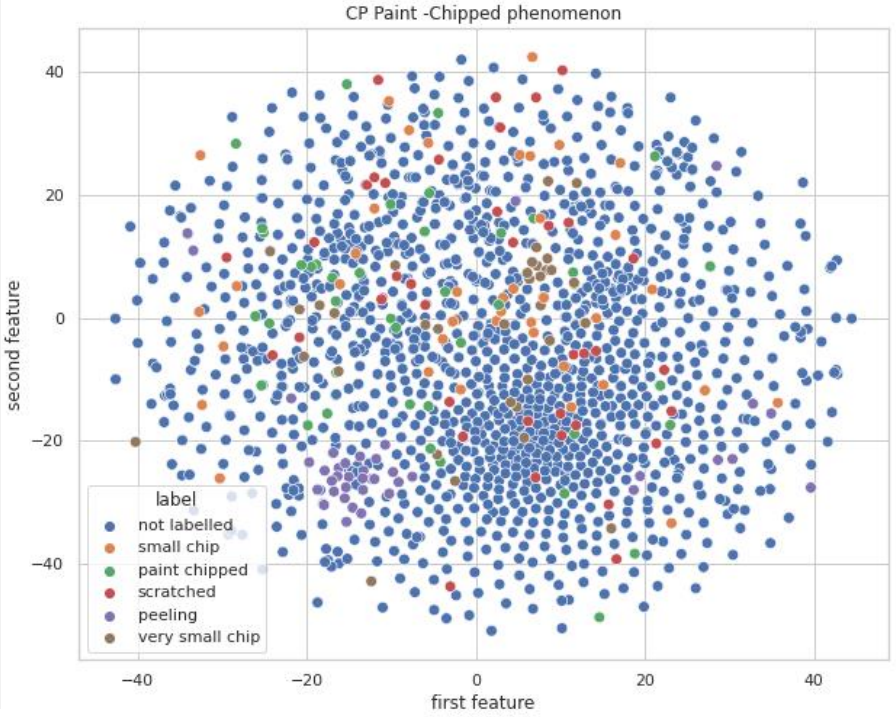


- After extracting final layer, we found [2,11,768] a 3-dimensional vector.
- We reshape the above vector to [2, 8448] a 2-dimesional vector for applying dimensional reduction method.
- And then applied dimensional reduction method **PCA**, we found below data points for both sentence for plotting.
  - Doc1: (-8079.903809　2.142040e-07)
  - Doc2: (8079.903809　2.142039e-07)

CP Paint –Chipped Annotation Visualization

Visualizing CP Paint -Chipped phenomenon using TF-IDF vectorization method

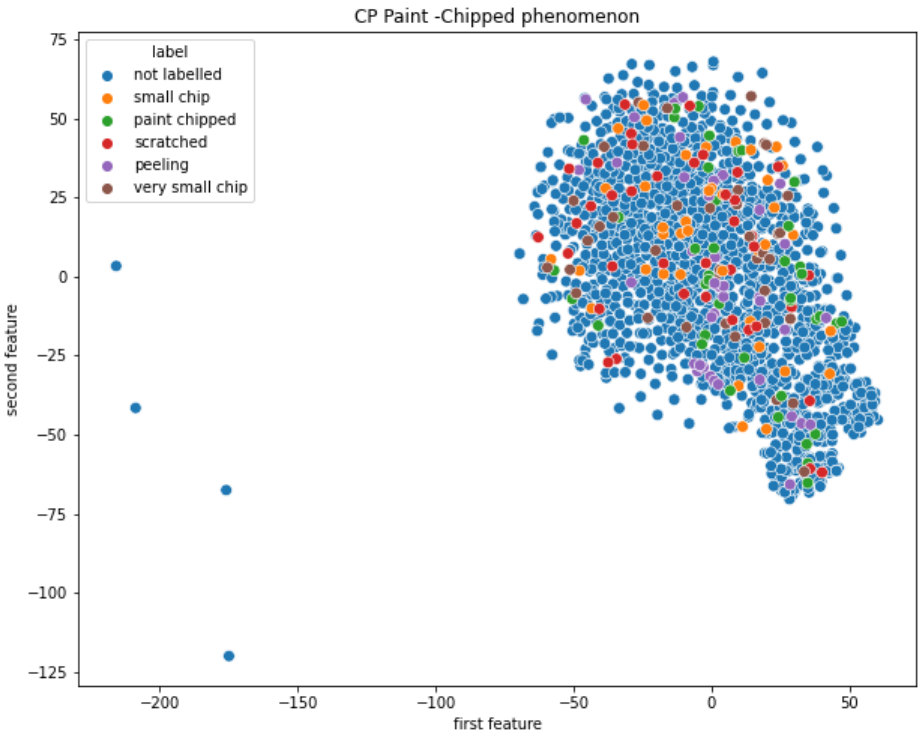
- As we can see, labels are covered most of the area of the data points.
- But again, label data points are scattered(overlap) like previous one.
- In CP phenomenon all five defined labels are generic(common for multiple items) labels. So, this is the reason behind overlap.



CP Paint –Chipped Annotation Visualization

Visualizing CP Paint -Chipped phenomenon using BERT vectorization method

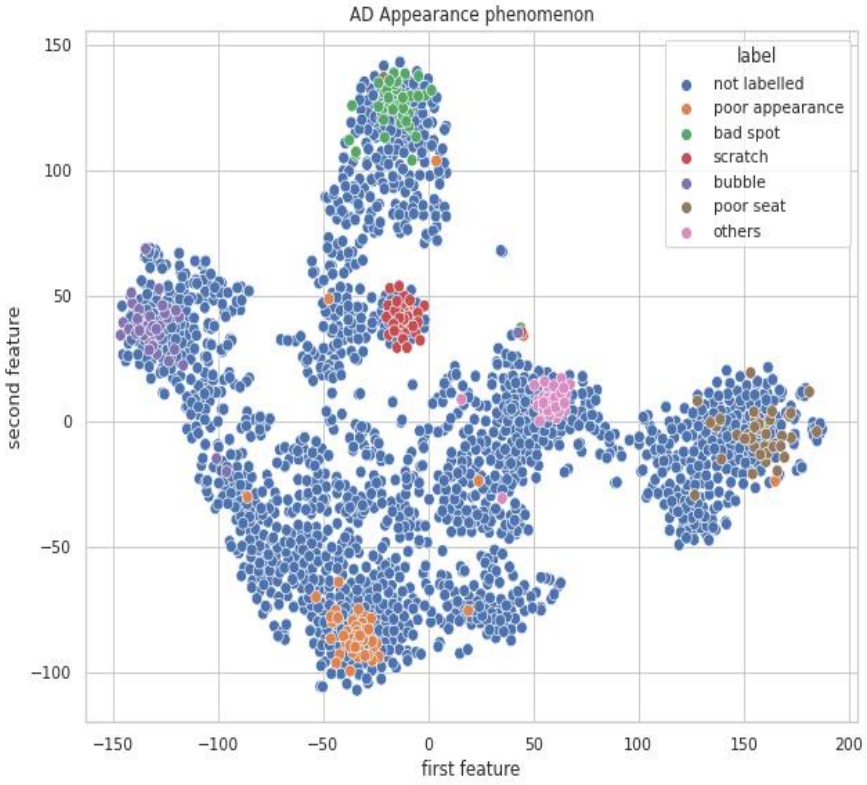
- Samples are more compressed compared to tf-idf vector. Defined labels are covered most of the area.
- Data points are more scattered compared to tf-idf. Overlapping problem occurs same as previous one.
- There are few anomaly in bert vectorization.



Fine-tuning BERT

After fine-tuning BERT model, visualizing the classified samples for AD Appearance phenomenon .

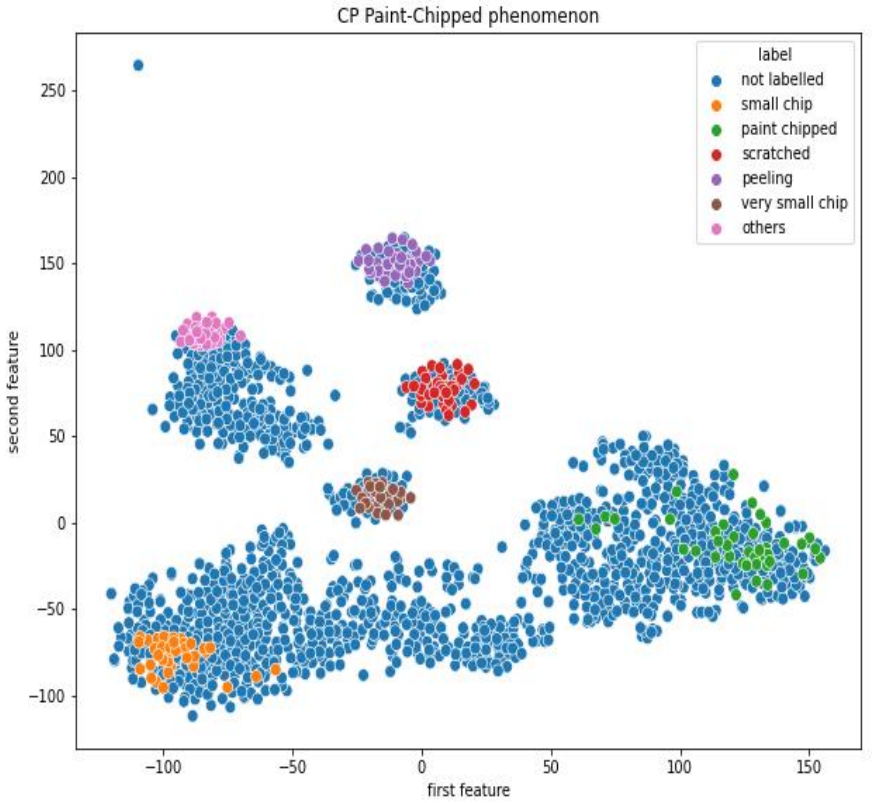
- Extracting the features for all the samples for AD Appearance phenomenon using BERT fine-tuned model trained by AD Appearance annotated datasets.
- As we can see the graph, BERT model's leaned well the data samples and data points are well classified amongst the classes.



Fine-tuning BERT

After fine-tuning BERT model, visualizing the classified samples for CP Paint Chipped phenomenon .

- Extracting the features for all the samples for CP Paint Chipped phenomenon using BERT fine-tuned model trained by CP Paint Chipped annotated datasets.
- As we can see the graph, BERT model's leaned well the data samples and the data points are well classified amongst the classes.



Thank you