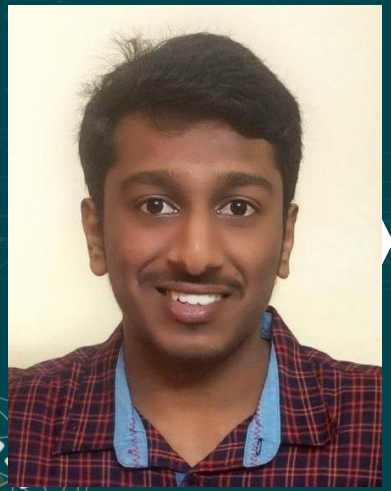
Amazon ML Challenge Finale

DeVaSh.Ai



Yerram Varun
IIT Guwahati

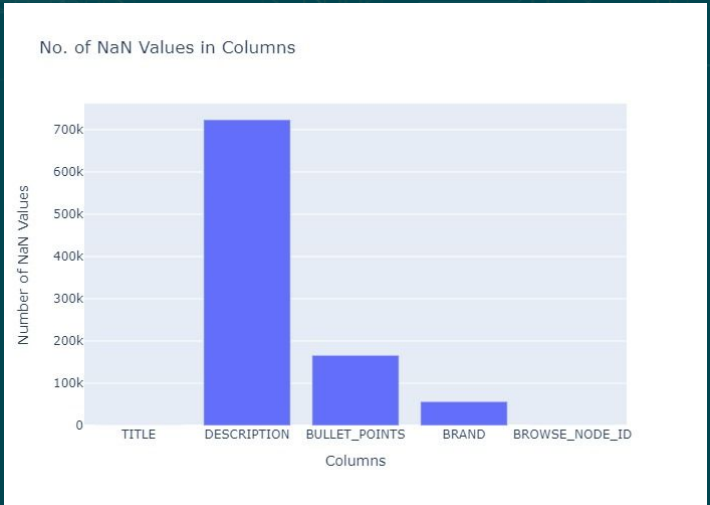
Debarshi Chanda
IIT Guwahati

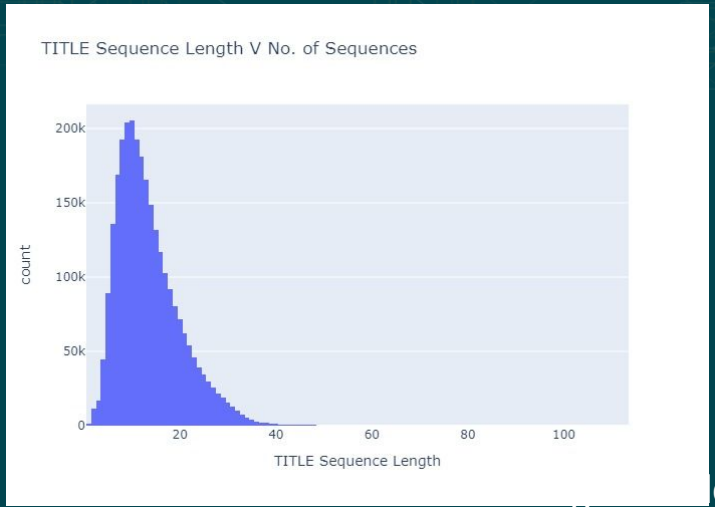
Shreya Sajal
IIT Guwahati

Aishik Rakshit
IIT Guwahati

OBSERVATIONS ON DATA

- The Dataset is huge (2.48 GB).
- TITLE has **least** number of NaN values and **smallest** sentence lengths.
- TITLE, DESCRIPTION and BULLET_POINTS have **redundant** information
- **3603** out of total **20302** brands available in TEST data are not available in TRAIN data.





Approach Overview

K-Nearest Neighbours + Embeddings

- **Arc-Face Embeddings** followed by **KNN** Classifier
- *TITLE* Feature used

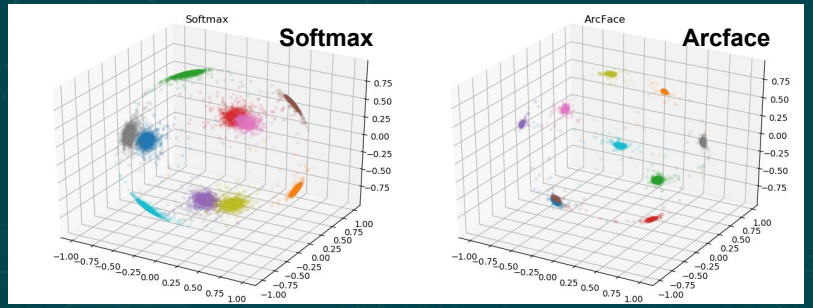
- Models trained using this recipe:
  - **XLM-RoBERTa**
  - **DistillBERT**
  - **XLNet**
  - **MPNet**
  - **TF-IDF**

Text Classification

- End-to-End Classification
- All Product features used

- Model Trained using this recipe:
  - **RoBERTa**
  - **BERT**
  - **XLNet**
  - **GPT-2**

TRAINING SPECIFIC DETAILS



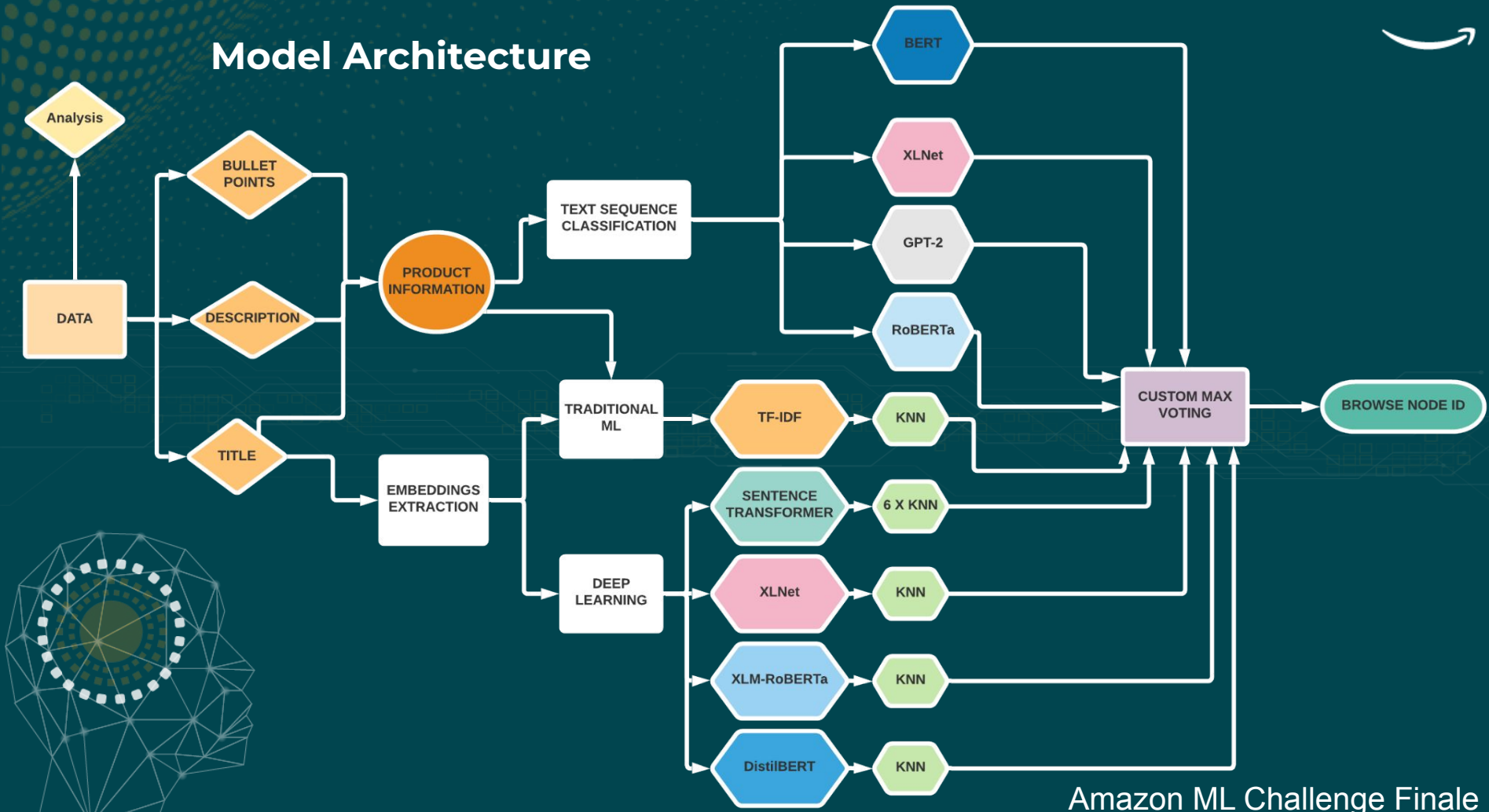
ArcFace ensures similar class embeddings to be close and dissimilar embeddings to be far.

K-Nearest Neighbours Classifier

- We applied KNN classifier on the embeddings we obtained from the Trained models.
- As we increased the data for training, we observed better results with lower K
- This way we give rare classes an opportunity to appear

Training is done on 15 Lakh rows on average for each of the models.

Model Architecture



Amazon ML Challenge Finale

Future Areas of Improvement

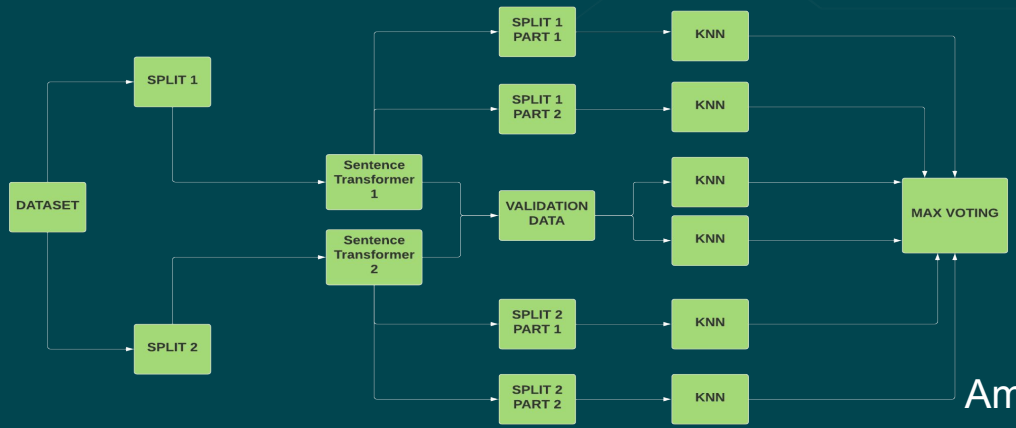
We can extract useful information from DESCRIPTION and BULLET_POINTS

A student model can be trained with the ensemble as Teacher model using Knowledge Distillation

Normalization and Stacking embeddings before applying KNN classifier

More Generalized models can be trained by using more data and training for longer duration

Alternative and Better approaches to KNN classifier can be explored, for example PCA on Embeddings, Kmeans++, XGBoost, etc



BEST SINGLE MODEL!