

Case Study – NLP Python Application

Context

- A global review and recommendation platform were looking to improve the efficiency of their matching and recommendation system when categorising similar items from e-commerce platforms.

Issue

- Data scrapped from e-commerce platforms and websites does not always have a clear definition and cannot be categorized by simple rules.
- The process in place included regex rules with specific categories to identify items which was prone to produce a high number of false negatives.

Approach

The Vox team was responsible for a wide range of activities, including:

- Implementing a pretrained BERT Model (NLP transformers model trained on GBs of data) to run after the existing rules rejected items
- Implementing a Python script that extracted data from a MongoDB database and then categorized the data accordingly using the BERT model
- Updating the data in the database with the new categories
- Implementing a percentage likelihood metric to correctly determine what the item was
- Establishing thresholds so that only items that had a high likelihood of matching to a specific item were selected

Results

- The solution cut down false negatives by 90% improving the number of items correctly categorized.
- Allowed the solution to be implemented on multiple websites without the need to create new rules and instances.