barion

Zoltán Balogh, PhD
Senior Data Scientist
zoltan.balogh@barion.com

How we use NLP at Barion

Budapest ML 2022

CONTENTS

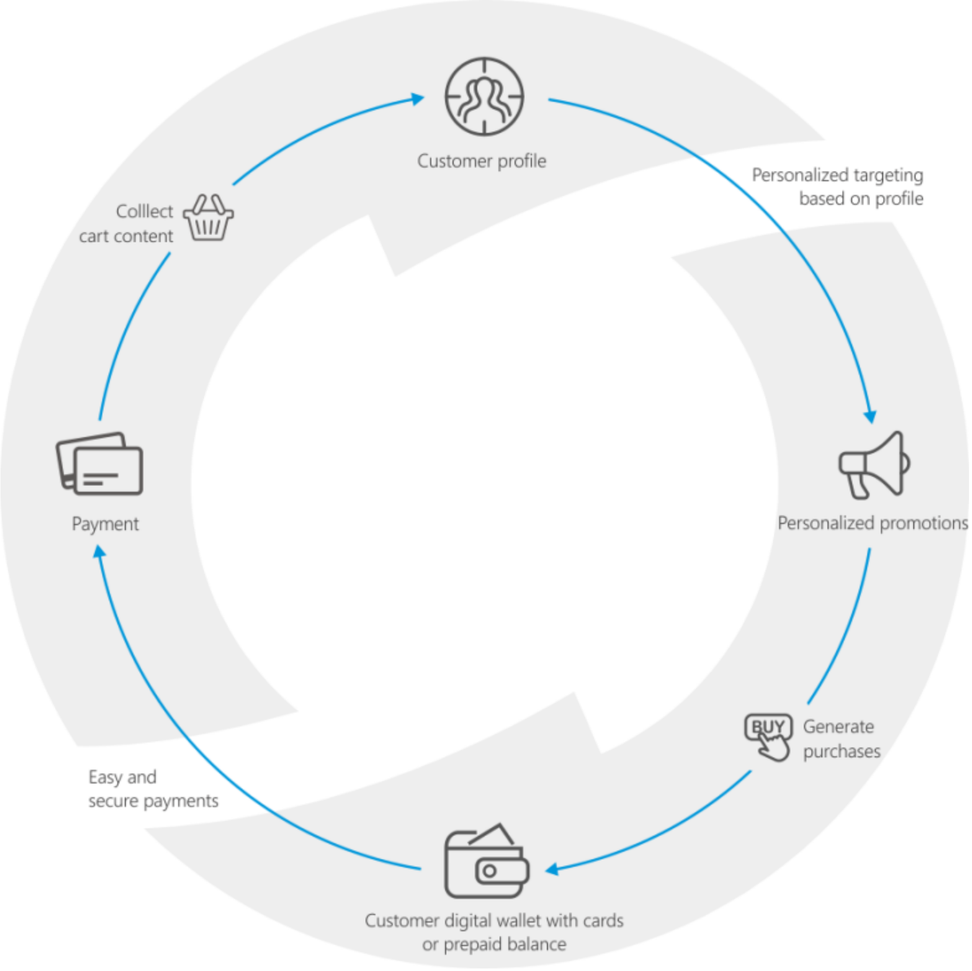
- About **Barion** & Data Monetization
- Problem statement
- Challanges of training set preparation
- Model building & training
- Used softwares

About Barion

- Licensed as an electronic money issuer
- With **Barion Smart Gateway** customers can easily and safely pay in more than 13 000 online shops mainly in CEE
- In exchange for **lower payment fees**, merchants can choose to **share consented data** of their **customers**, which is collected and stored in our data lake

Data Monetization

- **Barion Pixel** is a Javascript snippet built into the **merchant's website**
- With the customers' prior consent, the **details** of their **shopping behaviour** is sent to Barion
- The collected **data** is then **transformed** into profiles
- The created profiles are utilized in different advertising campaigns to enhance their targeting options
- When the customers visit webpages, **personalized promotions** are **shown** to them



Properties of Collected Data

- BarionPixel collects different events to capture the customers' shopping behaviour from more than 2000 merchants
- In this presentation we will be focusing on the contents of their shopping basket
- The items of the customers' basket are individually sent to and evaluated by the neural network
- The raw input of the neural network is the **category** and the **name** of the product or service provided by the shop

Requirements of the Output

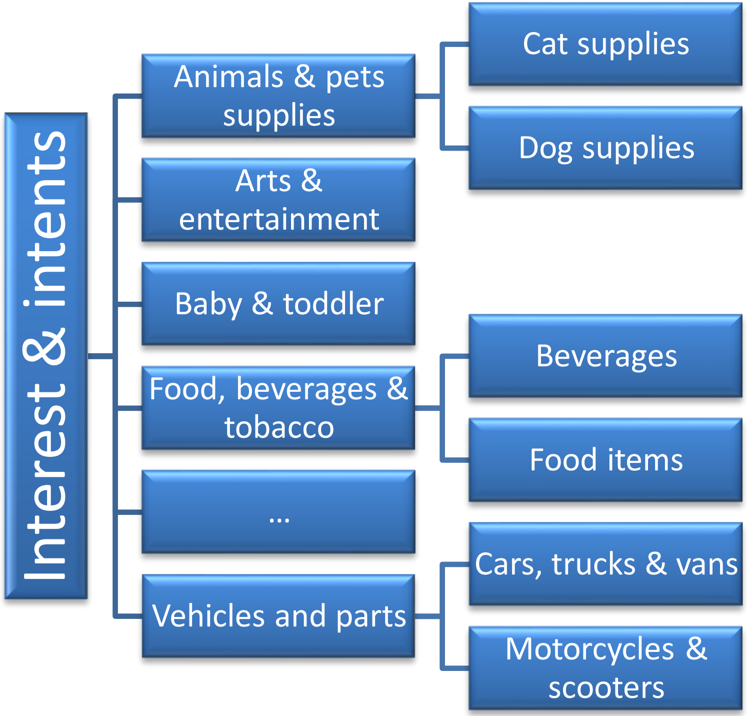
- Simple keyword search is not accurate and resource efficient
- A quick, robust and resource efficient model is needed that achieves high accuracy

- The output marketing segment should be produced from the model output with adjustable granularity
    - Ranging
        - From **getting acquainted** with the **product** (visited the product only once)
        - To the customers who **bought** the **product** or **have a long history of interest** for it
    - Accuracy/Confidence property

Segmentation Bases

- Marketing segmentation can be based on
  - Geolocation
  - Technographic
  - Demographic
  - **Interest and Intent**

- Models
  - **Neural network**
  - Regex



Interest & intents

Animals & pets supplies

Cat supplies

Dog supplies

Arts & entertainment

Baby & toddler

Food, beverages & tobacco

Beverages

Food items

…

Vehicles and parts

Cars, trucks & vans

Motorcycles & scooters

Challanges of NLP training set preparation

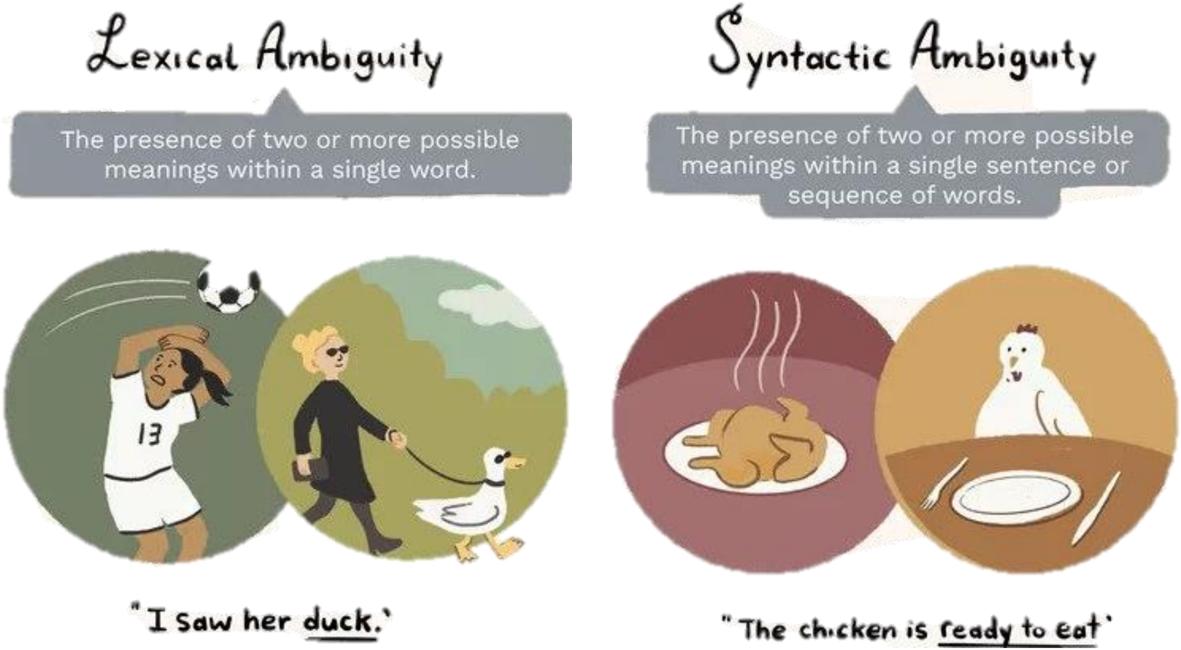

Lexical Ambiguity
The presence of two or more possible meanings within a single word.
13
"I saw her duck."
Syntactic Ambiguity
The presence of two or more possible meanings within a single sentence or sequence of words.
"The chicken is ready to eat"

Image source: https://www.thoughtco.com/ambiguity-language-1692388

Peek into the raw input data

- **Audi A6 autó izzó**

Disproportionate keywords within sample

- Újszülött pelenkák - Pop-in - Gyártók - Zöld Úton - Mosható
- Férfi MTB kerékpár - MTB kerékpárok - Kerékpárok - 5 - Bicaj
- AlphaOne DZ Inteligentní hodinky, bílé
- **Ledvance Planon Plus 30W 2700-6500K 595x595mm felületre szerelhető LED panel távirányítóval**

Noisy inputs

- **Női** felsők, **klassz** pólók hölgyeknek **kedvező áron**
- Line Sugar Effect Gel Silcare - Nail4U
- **Canon PG545XL**

Keyword not present

- **Asus P8H61-M LX2 1155 alaplap+ CPU hűtő**

Keywords might belong to a different category

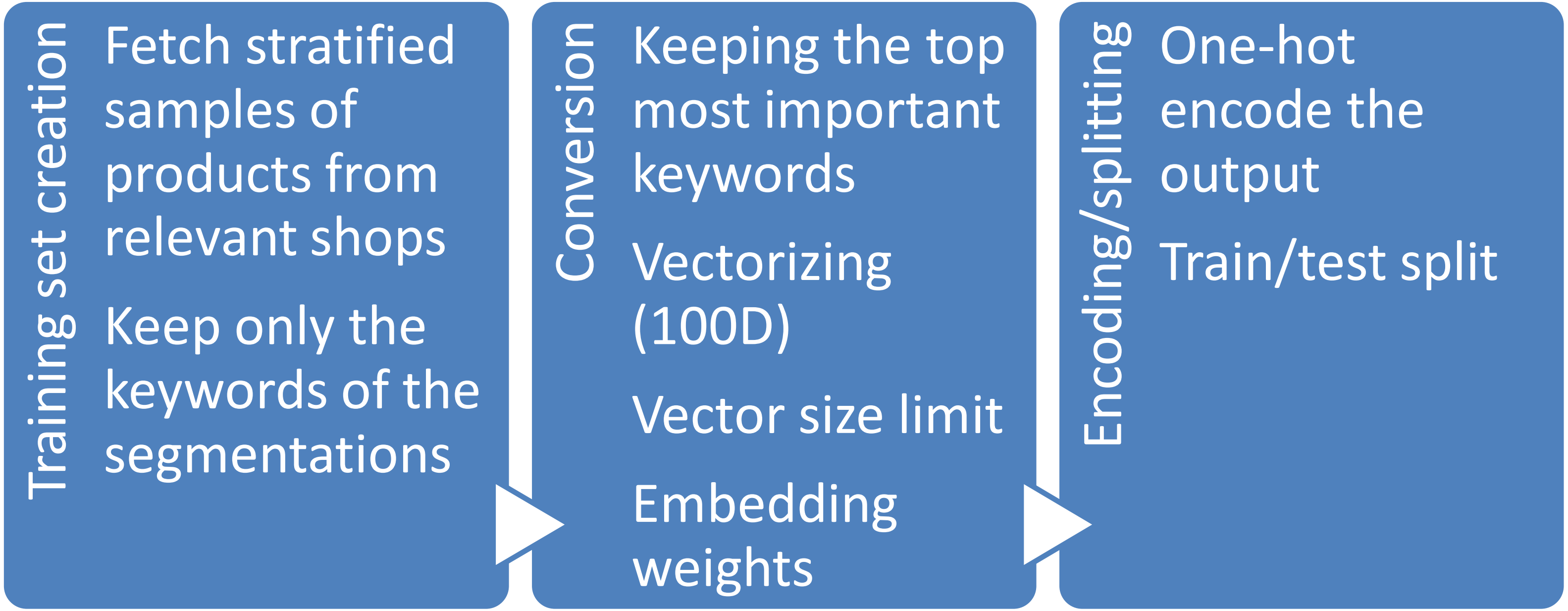
- Street Surfing Ripper Roller - Bloody Gold
- **;Névre szóló baba- Angyalka (Új) Karácsonyi limitált kiadás**

Category not clear

- **R15 Gyűrű**
- **CN-HG 95 lánc**

Ambiguous

Process of the preparation of the input



Training set creation

- Fetch stratified samples of products from relevant shops
- Keep only the keywords of the segmentations

Conversion

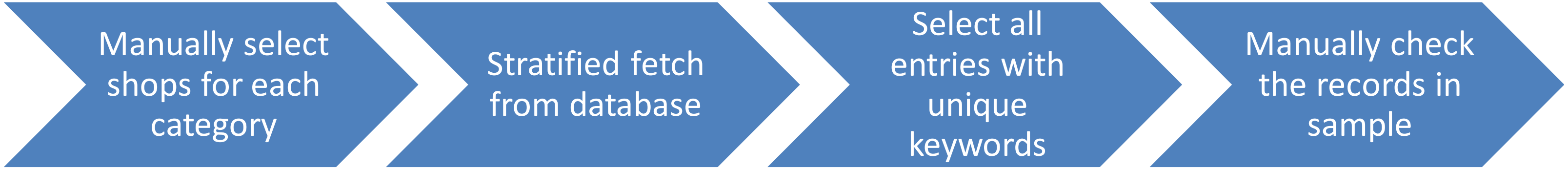
- Keeping the top most important keywords
- Vectorizing (100D)
- Vector size limit
- Embedding weights

Encoding/splitting

- One-hot encode the output
- Train/test split

Sampling

- Not all the categories have the same amount of quality entries
    - Stratified by the category of product/service
- The categories with excessive number of items need to be undersampled
    - Ensuring the include all the unique keywords relevant to the category



Manually select shops for each category

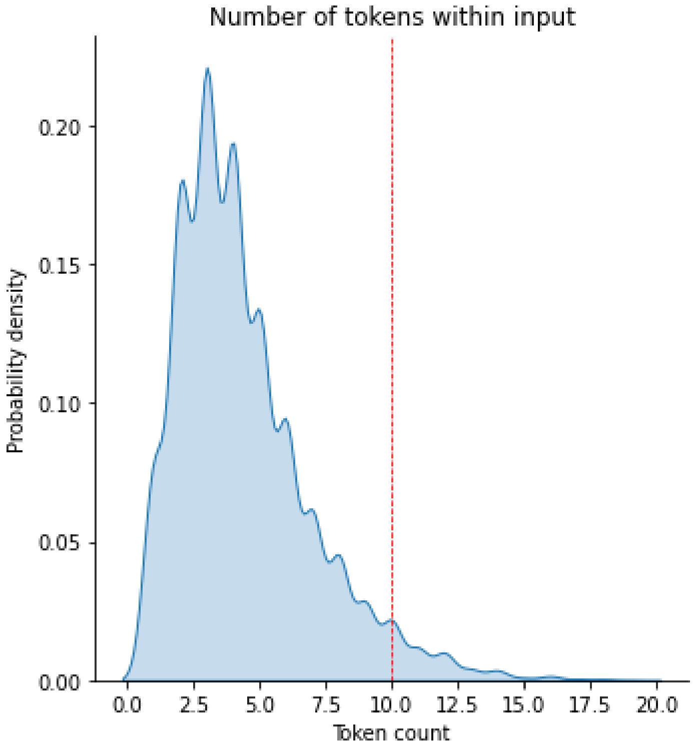
Stratified fetch from database

Select all entries with unique keywords

Manually check the records in sample

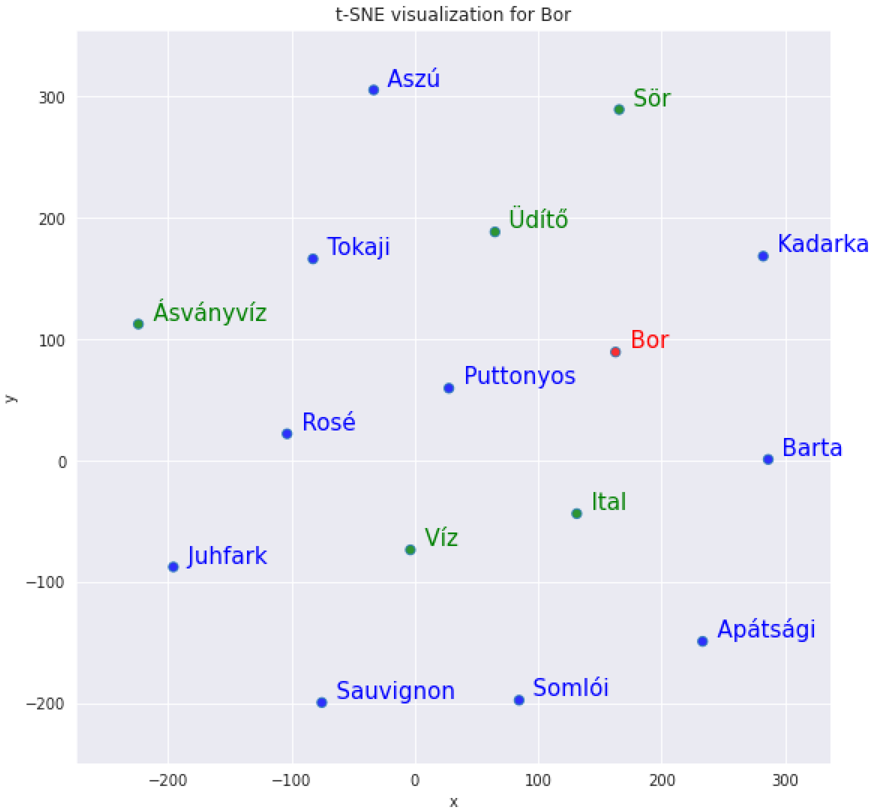
Token Cleansing

- Only the first 10 tokens within the input text is kept (as only 3.02% input entries are longer that 10 tokens)
- Stopword removal (tailored to the Hungarian market)
    - Colors (grey, yellow, green etc)
    - Measurement units (pcs, liter, xl, month)
    - Conjunctions (and, or)
    - Others (new, compatible, super, import, export, action, premium etc.)



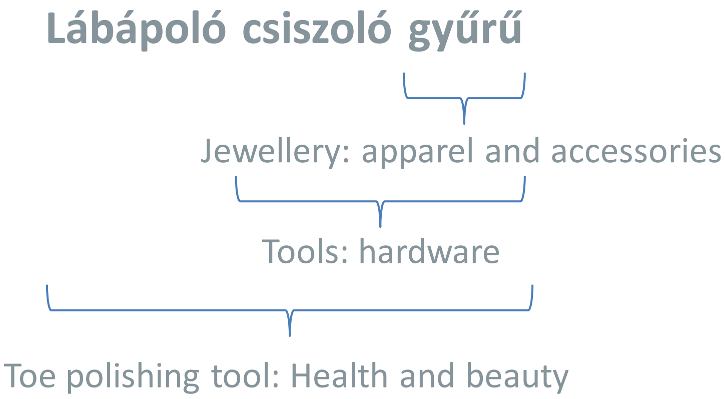
Vectorization

- ○ Vectorizing using gensim.Word2Vec
- ○ Converts all tokens to a 100 dimensional vector space
- ○ Ensures that words with similar meaning are closer to each other in space



Model architecture

- Requirements
  - The model should output only 1 category (no overlaps)
    - Binary crossentropy loss function
  - Needs to capture the word order in phrases effectively
- Tried different models of 2/3 layers
  - LSTM with heavy regularization and dropout
  - CNN
  - LSTM bidirectional

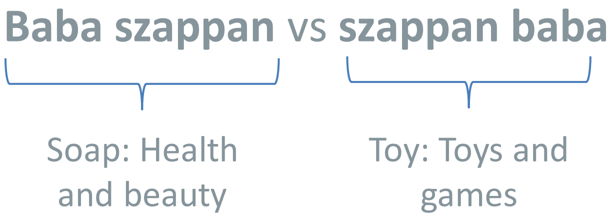


Lábápoló csiszoló gyűrű

Jewellery: apparel and accessories

Tools: hardware

Toe polishing tool: Health and beauty



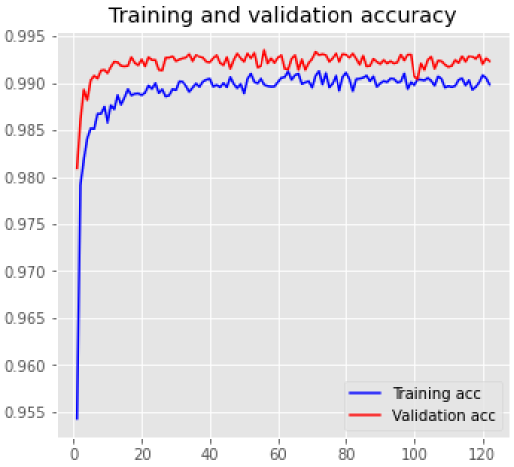
Baba szappan vs **szappan baba**

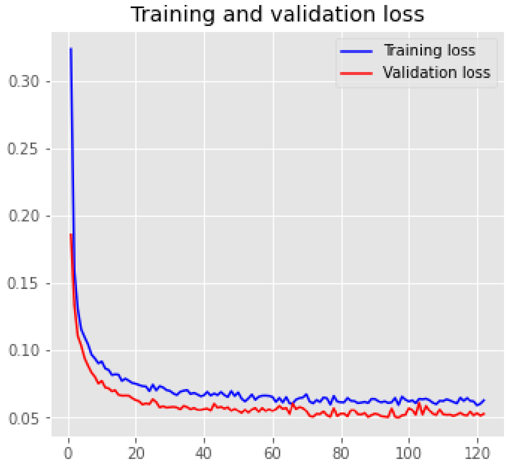
Soap: Health and beauty

Toy: Toys and games

Training & Output

- The output of the model and other features are saved into database linked to the user
- The output is not necessarily sales-ready
  - The database needs to be queried for a specific segment
  - The output can be customized further on different business demands





Different Architectures

- Sequential model
- First embedding layer with the weights of the pre-trained 100 dimensional embedding matrix

- Dense layer with sigmoid activation
- Adam optimizer, binary crossentropy loss function, optimize for accuracy

	LSTM	CNN
Number of layers	2-3	
Regularization	L1, dropout	
Filters	256, 128, (64/32)	
Extra		Global max pooling

The 2-layer LSTM performed the best

Result

- In recent a campaign for a **well-known multinational electronics company** promoting their child-care product-line, we provided our **„Baby and toddler”** category
- The targeted audiences in the given campaign were provided by **Google** and **Barion**
- The Barion segments outputted by the neural network gained **~25%** better performance, than the audiences provided by Google

Softwares, we use

python™

Spark

jupyter

Apache Airflow

K Keras

TensorFlow

barion

Thank you for your attention!

Zoltán Balogh, PhD

Senior Data Scientist

zoltan.balogh@barion.com

www.barion.com