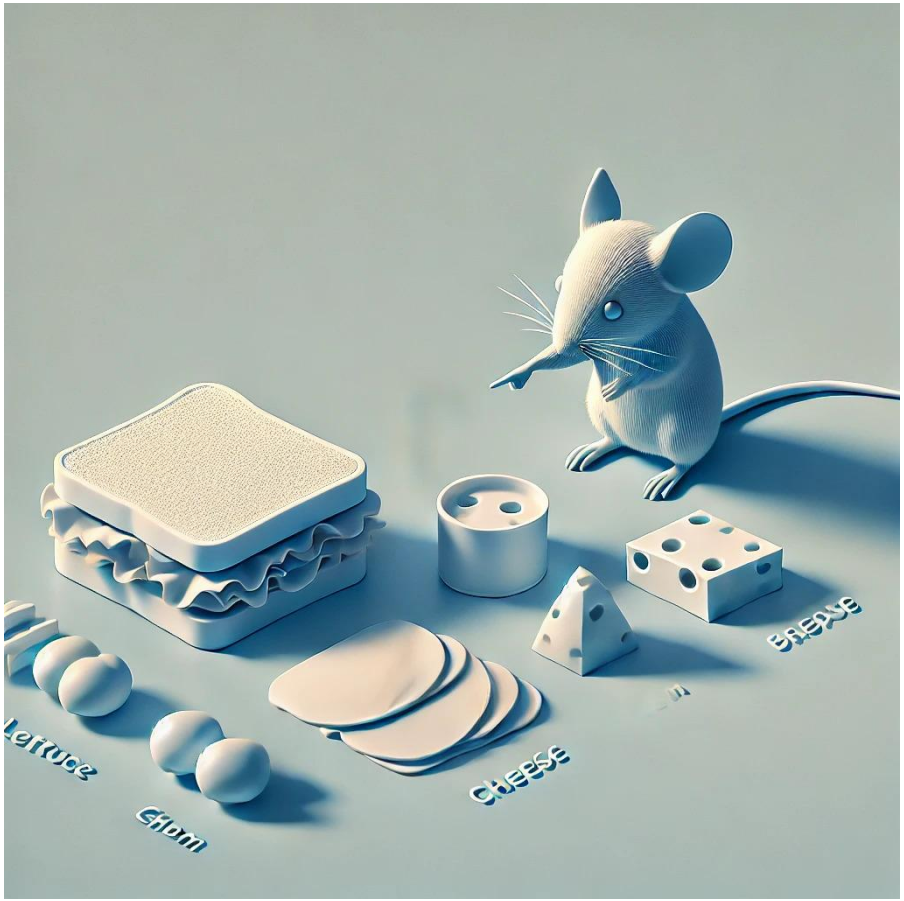
# Iron Rat race

LUIS

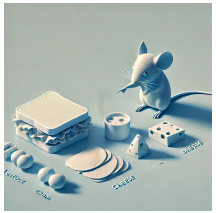

LADDER


Cheese

RAW DATA

SENTIMENT

CLUSTERS



CLEAN DATA

SUMMARIZER

REVIEWS

# DATA PREPROCESSING

# DATA PREPROCESSING

- Small dataset (1429.csv)
- Transformers: robust to noise
- No need to lowercase or remove punctuation
- Text without redundant special characters
- Concatenation of **title** and **text** columns
- Removed NaNs

# SENTIMENT ANALYSIS

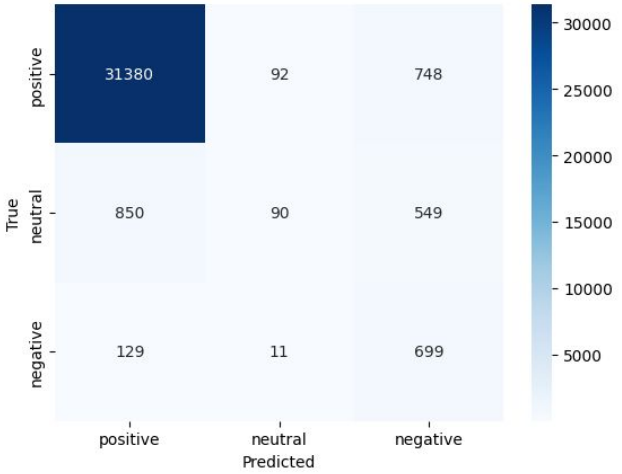
# SENTIMENT ANALYSIS

- **Star-rating analysis**
  - Assigned **sentiment** to stars (4 & 5 - positive, 3 - neutral, 1 & 2 - negative)
  - **Imbalanced**: positive ratings overrepresented (>90%)
  - Tried to generate **synthetic** reviews: didn´t work
  - Finally did not under- or oversample: could affect the clusterizer and summarizer

- **Customer rating not fully reliable**
  - **Zero-shot** classification with *facebook/bart-large-mnli*
  - Analized results and cleaned errors (0.32%)

# SENTIMENT ANALYSIS

- **Evaluation**
  - **Accuracy**: 0.93%
  - **Precision:** 0.93%
  - **Recall:** 0.93%
  - **F1 Score**: 0.92%



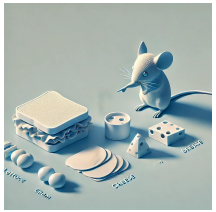

- **Confusion Matrix**
  - Perform best on the positive class
  - Neutral class most challenging to classify
  - Good at identifying negative instances, with some confusion with positives

# SENTIMENT ANALYSIS

- First attempt with *distilbert-base-uncased-finetuned-sst-2-english*

- **Saved** csv file with predicted labels

# CLUSTERING

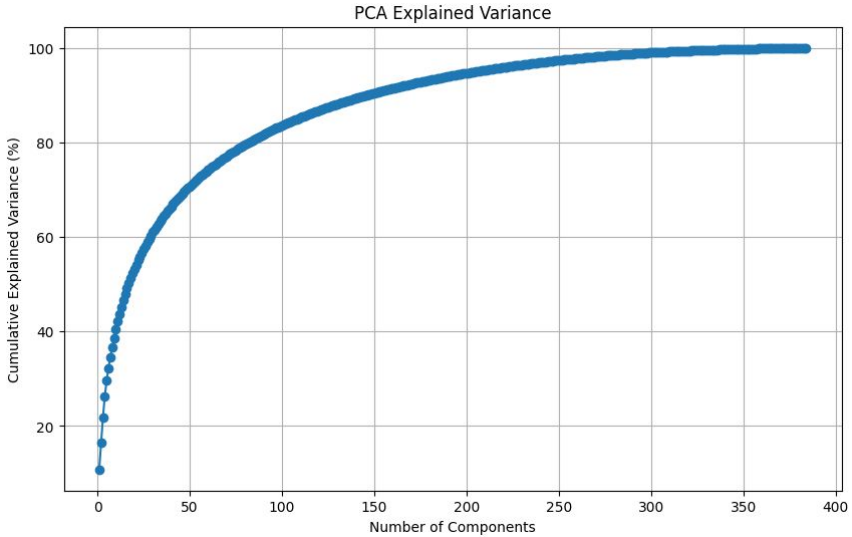


# CLUSTERING

- **Embed** text with Transformers (**sentence-transformers/all-MiniLM-L6-v2**)
  - Text to numeric representations & capturing semantic meaning
- **Cosine** similarity
  - Measures the cosine of the angle between two vectors
- Dimensionality Reduction (**PCA)**
  - Reduces dimensionality while preserving variance
  - Efficient, preserves global structure

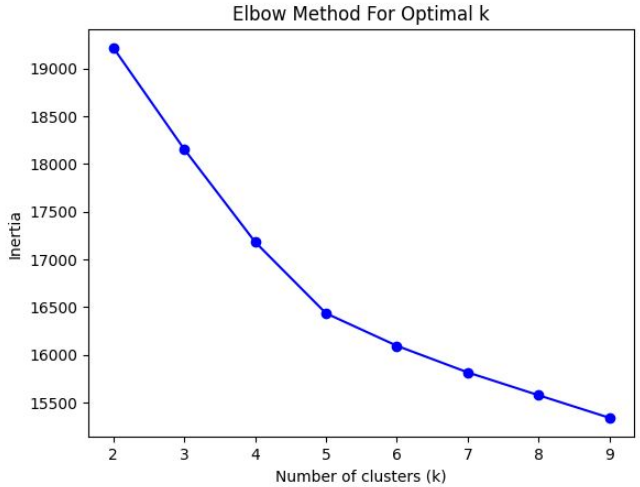
# CLUSTERING

- **Cumulative** explained variance
  - Obtain number of components for PCA (**125**)

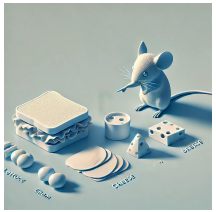


# CLUSTERING

- **K-Means**
  - Elbow Method for optimal number of clusters (5)

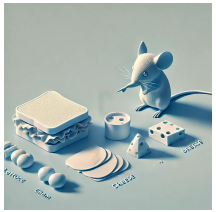


# CLUSTERING



- **K-Means**
  - Analise cluster content
  - Identify dominant product categories
  - Common words per cluster

# CLUSTERING



**Cluster 0 = Tablet** (Tablet)

**Cluster 1 = Smart Speaker** (Echo, Alexa)

**Cluster 2 = E-book** (Kindle)

**Cluster 3 = Smart TV** (TV, Fire)

**Cluster 4 = Generic Love** (Great, Product, Love)

Most repeated word = **GREAT** 🤔

# CLUSTERING

- **Assign labels to cluster**

- **Silhouette Coefficient (cluster dispersion) = 0.10**
  - Between -1 and 1
  - Not misclassified but neighbouring clusters nearby

- **DB Index (avg similarity of each cluster and its most similar one) = 2.67**
  - Between 0 and 1 = better clustering
  - Suggest poor clustering

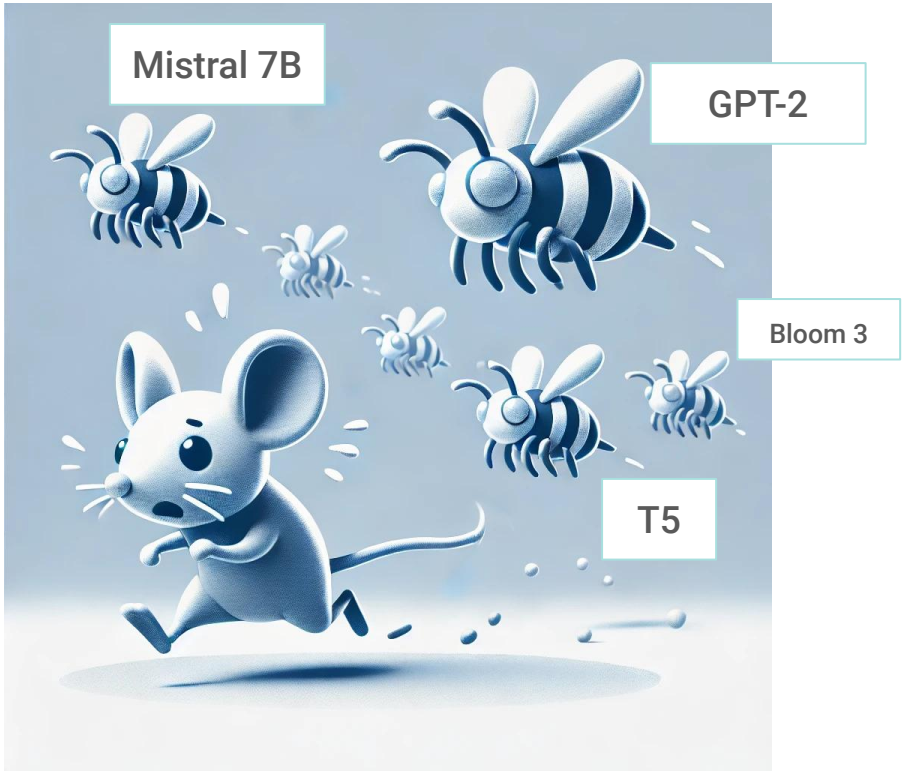
# CLUSTERING

- **Removed words with Regex to improve metrics**
  - Great, product, love
  - Same clusters, no improvement
- **1st try clustering with product title**
  - Loses the point of the task

# SUMMARIZER

# SUMMARIZER

# SUMMARIZER



# SUMMARIZER

- **Removed** the “Generic Love” category & **concatenated** reviews + cluster + sentiment

- **Mistral 7B, T5, Bloom 3**

  - Didn´t manage to fully implement them

- **GPT-2**

  - High training (2) and validation (2.31) loss
  - Fine tuning: training arguments & prompt
  - None acceptable results

# RECAP & TAKEAWAYS

- **Small dataset**
- **Sentiment** with *facebook/bart-large-mnli*
- **Clustering** with *sentence-transformers/all-MiniLM-L6-v2* and K-Means
- **Summarizer** didn´t success (*GPT-2, T5, Mistral 7b, Bloom 3*)

# RECAP & TAKEAWAYS

- **Small dataset**
- **Sentiment** with *facebook/bart-large-mnli*
- **Clustering** with *sentence-transformers/all-MiniLM-L6-v2* and K-Means
- **Summarizer** didn´t success (*GPT-2, T5, Mistral 7b, Bloom 3*)

- **Find the balance between research and implement/test/fine-tune models**
- **Clean and balance the data as much as possible**

# RECAP & TAKEAWAYS

- **Small dataset**
- **Sentiment** with *facebook/bart-large-mnli*
- **Clustering** with *sentence-transformers/all-MiniLM-L6-v2* and K-Means
- **Summarizer** didn´t success (*GPT-2, T5, Mistral 7b, Bloom 3*)

- **Find the balance between research and implement/test/fine-tune models**
- **Clean and balance the data as much as possible**

https://github.com/luislague/IronRat