From Classic to Cutting Edge Text Classification:

Generating Customers Insights with Topic Modelling & Hugging-Face SetFit Method

Sultan Al Awar
PyData London 2024

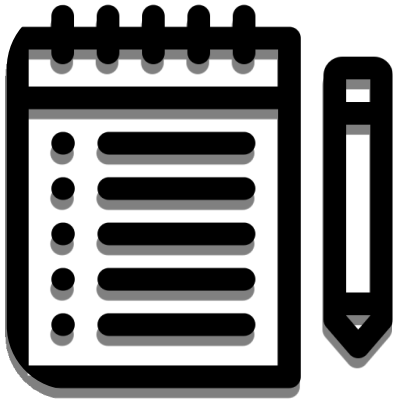
Friday, June 14
Leonardo Royal Hotel - Minories Suite

About me

- Data and ML enthusiast
- **Data Scientist** at E.ON Next
- **Certified AWS ML Practitioner** and **Subject Matter Expert** (SME)
- Latest DS products focus on Gen-AI call tagging, marketing propensity modelling, customers reviews classification, demand & supply forecasting, customer segmentation, recommender system, and customer lifetime value
- **MSc in Business Analytics** from University College London (UCL)
- Previous **Consultancy** experience
- Hobbies: Tag Rugby and Hiking

Agenda

- Text Analytics and Customer Experience
- Classic topic modeling, covering text pre-processing, feature engineering, and the LDA algorithm.
- Hands-on application - Unsupervised topic modeling
- Hugging Face SetFit few-shot learning Method
- Hands-on application - Fine-tuning a sentence transformer and generating model inference
- Wrap-up and Q&A



slido

What techniques and tools have you used or been exposed to for analysing customers reviews data?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

slido

What will be the business benefit of analyzing and classifying customer reviews?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Text Analytics in the Real World

Definition

- Text Analytics, is the process of transforming **unstructured data** to uncover meaningful information, patterns, and insights.
- It leverages techniques such as Generative AI and Natural Language Processing (NLP).

Data Types

- According to IBM, unstructured data comprises over **80% of all enterprise data.**
- Text data could take the forms of **survey responses**, **Trustpilot reviews**, **call center notes, product feedback, social media posts.**

Techniques

- **Sentiment Analysis**
- **Topic Modeling**
- **Named Entity Recognition**
- **Text Classification**
- **Information Retrieval**
- **Summarization**

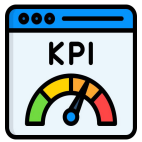
Boosting Customer Experience with Text Analytics

Key Benefits:

1. **Data-driven Customer Understanding and Insights:** reveals customers needs, preferences, trends, and pain points to make decisions accordingly.
2. **Efficient Data Processing:** automates analysis of large text data, saving time and resources.

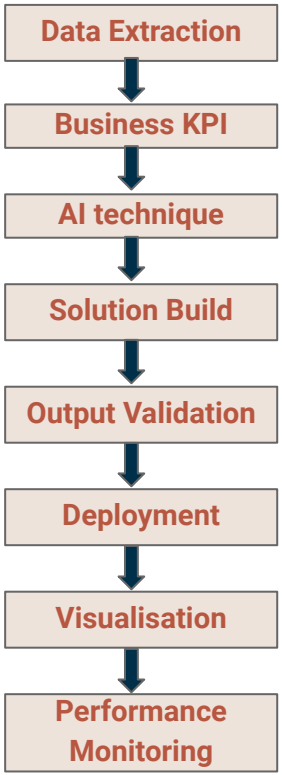
Key Performance Indicators:

1. **Loyalty Score Improvement:** Measure the change in Net Promoter Score, Customer Happiness.
2. **Churn Rate Reduction:** Measure the decrease in customer attrition rate.
3. **Revenue from Cross-Sell/Up-Sell**: Measure additional revenue from upselling or cross-selling to customers reviewing about loyalty programs, rewards, or incentives.



https://feedier.com/blog/text-analysis-the-benefits-of-analyzing-customer-feedback/

End-to-end Data Science Customers Reviews Project Approach



- Determine sources of text data like review sites (Trustpilot), customer feedback forms, social media
- Collect data using web scraping, APIs, or database queries and then conduct data cleansing

- Clearly define the business goal of the project
- Identify key performance indicators that will measure the impact derived like customer satisfaction, brand sentiment, churn rate

- Choose appropriate NLP technique such as topic modelling, text classification, summarization, etc.
- Decide on the tools and libraries to be used (e.g., NLTK, SpaCy, Gensim, Scikit-Learn, HF Transformers, LLMs).

- Extract relevant and recurrent topics
- Classify Reviews into predefined categories

- Use ML performance measures such as perplexity (topic modelling) or accuracy (text classification)
- Involve domain experts to review the results

- Productionise the model
- Make decisions on Batch versus Real-time inference

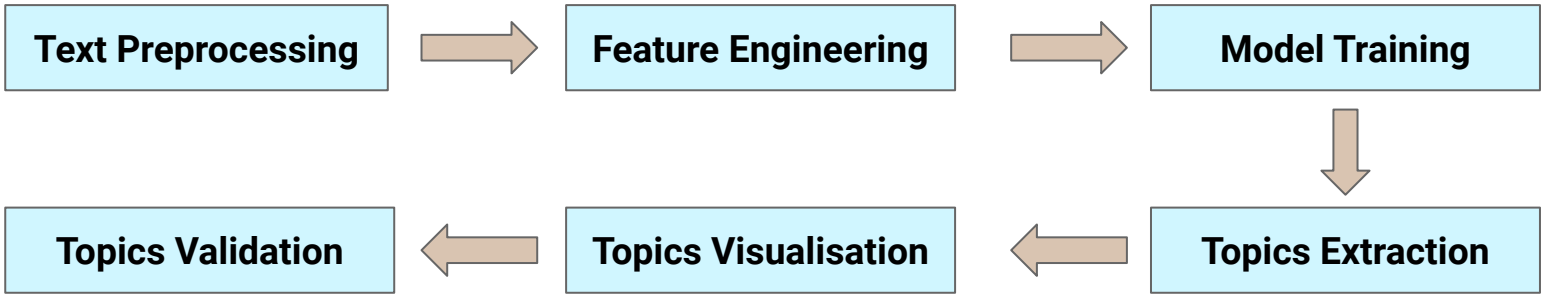
- Bring data to end users using BI platforms like Tableau or Microsoft Power BI through interactive dashboards

- Set up continuous monitoring to track the performance of the model
- Periodically evaluate the impact of the insights on business metrics and integrate stakeholders' feedback

Uncover Business Themes in Customer Reviews using Topic Modelling

- **Topic modeling** is an unsupervised learning technique used to identify and extract underlying themes or topics from a collection of unstructured customer reviews.

- For example, if utility company customer says *"The dual tariff is expensive"* while another says *"The new energy package is unaffordable"*, while the words they're using are different ('tariff' vs 'package') they are both referring to the same topic, and it could be categorized under *"Tariff Price"* theme.



Text Pre-processing: Key Preliminary Step before LDA Training

Remove words containing numbers and special characters

Python method: **re.sub** from Regular expression (re) library

Tokenize: break text into tokens

Python method: **word_tokenize** from nltk library

Lemmatize: obtain word origin within context

Example: running —> run; better —> good

Python method: **WordNetLemmatizer** from nltk library

Remove stop words

Python method: **stopwords list** from nltk library

Remove punctuations

Python method: **string.punctuation** from string library

Lower text and remove short words

Python method: **text.lower** and **len(text)**

Feature Engineering: Text Representation

Transforming tokens into numeric word embeddings

Doc2bow (Document to Bag-of-Words) Embeddings

- **Use a dictionary of all unique words** (vocabulary) across the corpus.
- **Count the number of occurrences** of each word appears in a document.
- **Represent the document as a list of tuples,** where each tuple contains a word's index in the vocabulary and its count.

➔ *Ignores word order and context*

TF-IDF (Term Frequency-Inverse Document Frequency) Embeddings

- **Term Frequency:** Measures ***how frequently a word*** appears in a document.
- **Inverse Document Frequency (IDF):** Measures ***how important a word*** is by accounting for its frequency across the entire corpus.
- This means that words that are more frequent in a document will have a higher weight, but if those words are also frequent in other documents, their weight will be reduced such as great, amazing,, etc.

➔ *Ignores word order and semantic meaning*

Most Popular Topic Modelling Algorithm: Latent Dirichlet Allocation (LDA)

Introduced in 2003

➔ **It aims at learning the topic distributions that best explain the content of the entire documents (corpus)**

➔ **Available Python libraries: Gensim & Scikit-learn**

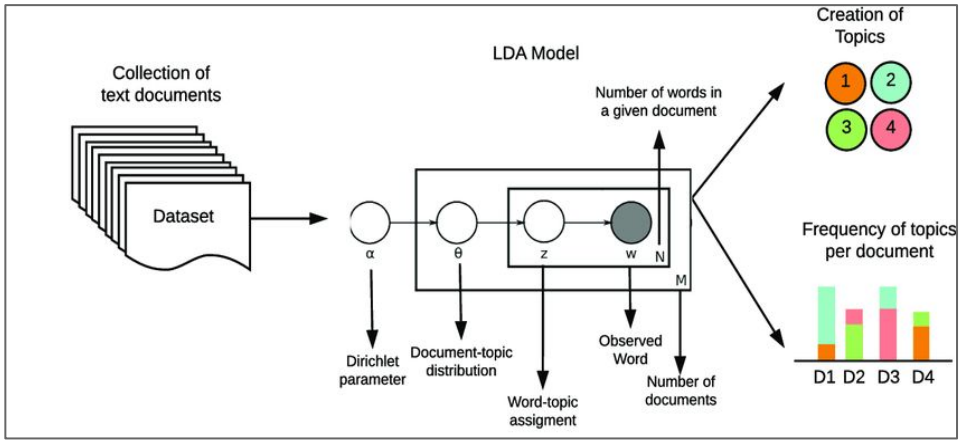


Image taken from: https://www.researchgate.net/figure/Schematic-of-LDA-algorithm_fig1_339368709

Generative Probabilistic Modeling Approach:

- **Identify topics distribution across documents**
- **Assign a word distribution to each topic**
- **Create topics and frequency of each topic per document**

Limitations:

- **Context**: Ignores the order and context of words within documents.
- **Interpretation:** Topics can sometimes be difficult to interpret meaningfully.

Hands-on Topic Modelling Application

Notebook: customers_reviews_topic_modelling.ipynb
Data: reviews.csv

Key Takeaways from the Classic Topic Modeling Section:

Topics Generation Challenges: 

- The success of topic generation depends heavily on **data quality and effective pre-processing techniques**. High levels of noise can hinder accurate topic identification.

Importance of Business Judgment: 

- Transforming clusters of words into meaningful, contextual themes requires **human insight and business understanding** to ensure relevance and applicability.

Project Progression: 

- Topic modeling is a critical step that facilitates the **transition to data labelling and text classification**.

SetFit: Efficient Few-Shot Learning Without Prompts

Introduced in 2022

It applies few-shot learning, where the model is given a few examples of the target task (what success looks like)

Two Step Modelling Approach:

- **Fine-tune a Sentence Transformer model on a small number of labeled examples (typically 8 or 16 per class).**

- **Train a classifier head on the embeddings generated from the fine-tuned Sentence Transformer.**

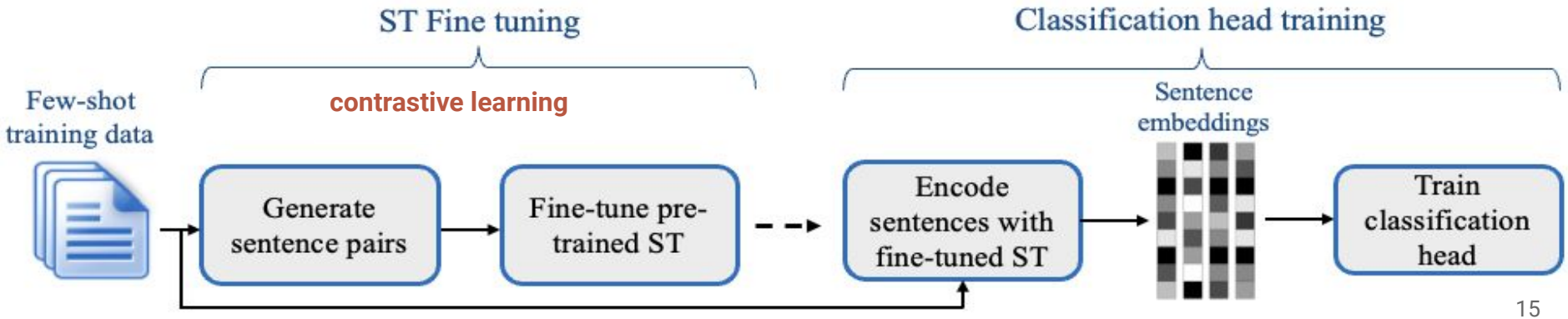


Image taken from: https://huggingface.co/blog/setfit

Key Advantages of Setfit

Efficiency in Few-Shot Learning:
- Requires only 8 to 16 labeled examples per class for fine-tuning

High Performance:
- Faster training and inference
- Achieves similar accuracy to larger models like GPT-3

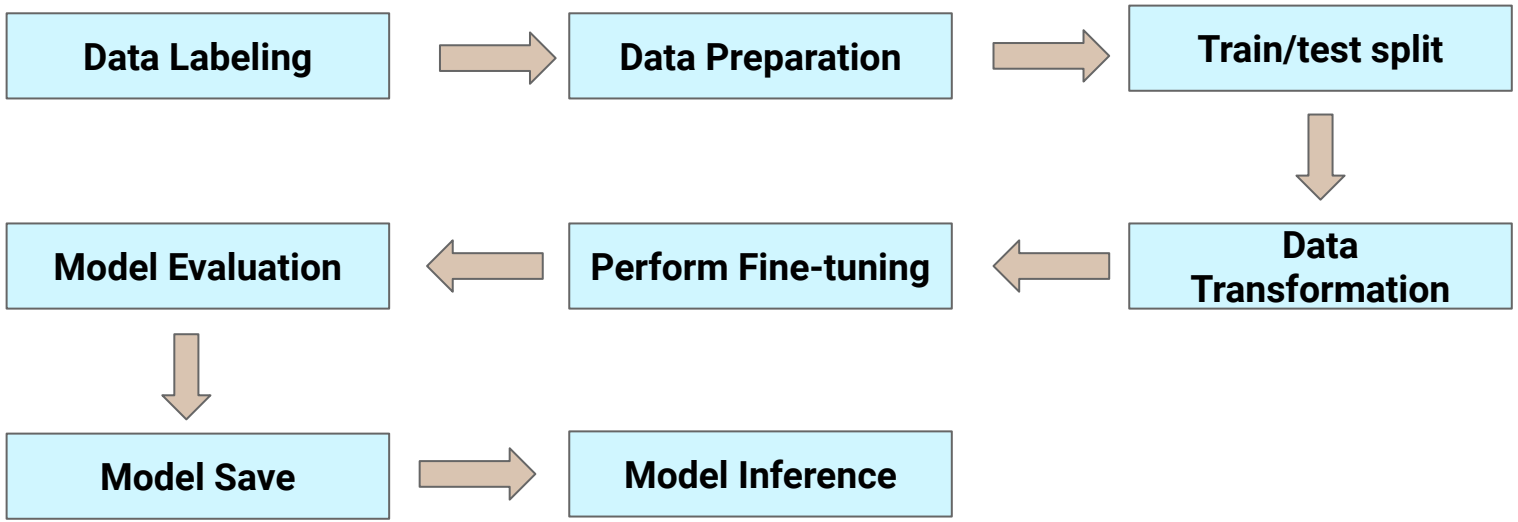
Prompt-Free Framework:
- Eliminates the need for handcrafted prompts
- Simplifies and streamlines the fine-tuning process

Multilingual Support:
- Can classify text in multiple languages with a multilingual checkpoint

Fine-tuning Setfit Implementation Steps

- **Setfit is available through the Hugging Face setfit library with the following classes:**

➔ SetFitModel: Loads the pre-trained sentence transformer
➔ Trainer: Handles the training process
➔ TrainingArguments: sets the training configurations



Generate Inference using Fine-tuned Setfit

Share the fine-tuned model to Hugging Face

Push the new fine-tuned to HuggingFace using either:

1) **Web interface**
2) **Client API - trainer.push_to_hub**("model-name")

Load the new customised fine-tuned model

fine_tuned_model = SetFitModel.from_pretrained(MODEL_ID)

Generate predictions on real (unseen) data points

fine_tuned_model.predict(data)

Model Evaluation & Performance Monitoring

Model training metrics

- Accuracy
- Recall
- Precision
- F-1 score

Baseline predictions using Python text matching techniques

Define Keywords for each Class/Category

↓

Search the text to detect the presence of these predefined keywords and phrases based on the assigned predicted theme

↓

Compare the baseline themes against the predicted themes

Python Libraries

FuzzyWuzzy:
It computes similarity ratios between text data and measures the transformation required to match string A to string B.

RegEx:
It searches a document for specific phrases or pattern.

Hands-on Application: Fine-tuning Setfit & Performance Monitoring

Notebook: customers_reviews_setfit.ipynb
Data: labelled_reviews.csv

Key Takeaways from the Set-fit Fine-tuning Section and Potential Next Steps:

Implement Robust Evaluation Mechanisms: 

- Develop a straightforward approach to generate baseline predictions. Compare these with the advanced model's output

Explore Current Gen-AI approaches: 

- Consider Large Language Model (LLM) prompt engineering or few shot learning capabilities for text classification with less technical complexity and overhead

Thank you for your attention.
Any questions.