

ML 2022 Text Classification Theme Group

InKyung Choi (UNECE)

Based on the theme group activities

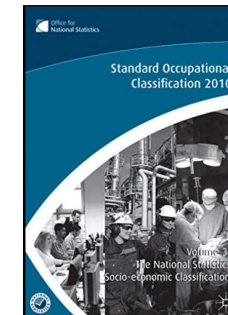
ML Group 2022 Webinar (November 30th, 2022)

Background

- Text classification tasks in the statistical organisations
 - Classifying job description **from survey** into SOC
 - Classifying economic activity **from register** into NACE
 - Classifying production description **from web** into ECOICOP
 - Classifying posts **from social media** into positive / negative sentiments



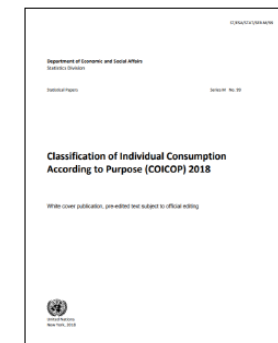
Survey response “I am a cook at an Italian restaurant”



SOC “35-2014.00
Cooks,
Restaurant”

Background


- Text classification tasks in the statistical organisations
 - Classifying job description **from survey** into SOC
 - Classifying economic activity **from register** into NACE
 - Classifying production description **from web** into ECOICOP
 - Classifying posts **from social media** into positive / negative sentiments



COICOP “01.2.2
Mineral waters,
soft drinks, fruit
and vegetable
juices”

Background

- Text classification tasks in the statistical organisations
 - Classifying job description **from survey** into SOC
 - Classifying economic activity **from register** into NACE
 - Classifying production description **from web** into ECOICOP
 - Classifying posts **from social media** into positive / negative sentiments
- Application in this area was the most popular in 2019-2021



ML can help
conduct task in a
more efficient way
and make use of
new data sources

Aim of Theme Group in 2022

Knowledge sharing and peer-review platform

ID	Month	Presentation Title	Speakers	Data	Methods
0	Mar	Kick-off meeting	-	-	-
1	April	Use of ML techniques for classification problems related to CPI	Vladimir G. Miranda, Lincoln T. da Silva (IBGE, Brazil)	Product description from web-scraped data	TF-IDF; naive bayes, logistic regression, SVC, SGD, Random Forest, XGBoost; LIME
2	May	Matching Big Data to Official Statistics Classifications	Alessandra Sozzi, Alberto Sanchez (IMF)	Google trends, Google places, Indeed job postings	direct matching, fuzzy matching, TF-IDF, Best Matching 25; Transformer for translation
3	June	Triaging Enquiries using Multilingual Transformers Model	Joanne Yoon, Alexandre Istrate, Shirin Roshanafshar (Statistics Canada)	Client enquiries	Multilingual BERT, XLM-MLM en-fr, XML-RoBERTa
4	Aug	Codification of firm activity from free text descriptions	Tom Seimandi (Insee, France)	Economic activity from business register	Fasttext, Softmax classifier
5	Sept	New model for coding using Deep Learning	Jael Perez, Alejandro Pimentel (INEGI, Mexico)	Economic activity from survey, Wikipedia text (for word embedding)	Fasttext, Bi-GRU, Softmax classifier
6	Oct	Unsupervised topic modeling and text classification using top2vec and lbl2vec	Michael Reusens (Statistics Flanders)	Company web pages	top2vec, lbl2vec
7	Nov	Wrap-up meeting	-	-	-

1. Use cases in statistical organisations

- Application of ML for classifying textual responses in survey questionnaire continues to be a solid use case.
- ML-based methods are particularly indispensable when it comes to big data which is becoming an increasingly important data source for statistical organization
- ML can help statistical organisations providing new services (e.g., policy makers, other government agencies)

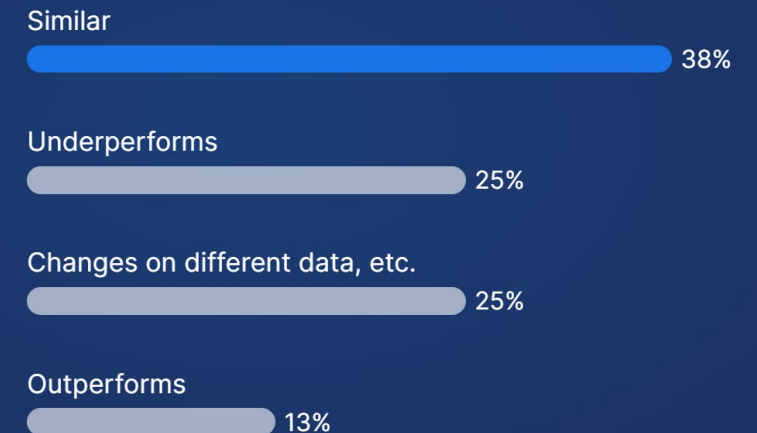
2. Is it worth changing to more advanced ML?

- NLP / ML is a fast-changing field

2. Is it worth changing to more advanced ML?

- NLP / ML is a fast-changing field
- Existing ML methods should be replaced by modern ones?

What were your experiences in using more sophisticated ML models compared to "classical" ML models?

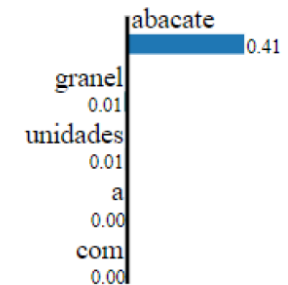


2. Is it worth changing to more advanced ML?

- NLP / ML is a fast-changing field
- Existing ML methods should be replaced by modern ones?
- Accuracy
 - Depending on the complexity of target texts
 - Be careful with what training data set is used
- Explainability
 - To address concerns around “black-box” models, explainers like LIME can be used

LIME (Local Interpretable Model-agnostic Explanations)

True: Abacate --> Pred: Abacate | Prob: 0.17
NOT Abacate Abacate



Text with highlighted words

abacate a granel com 2 unidades

3. How to address class imbalance?

- In the classification, the prediction performance is often poorer for rare classes compared to more prevalent classes as there might not be enough data for the ML model to learn for the rare classes.
- Several strategies were observed: oversampling, adapting threshold depending on the uncertainty of prediction, augmenting data using auxiliary data set (e.g., add CPI description)
- It is difficult in general to have a model that has good performance for every class when there are many classes in the data. It would be advisable therefore first to curate the data set beforehand

4. Beyond Proof of Concept (PoC)

- Even with proven validity, many challenges to turn the solution into a real solution in production
- Consider the maintenance strategy
 - For example, the confidence levels of ML predictions are used as a basis to select data to be manually labelled (i.e., data corresponding to low confidence levels are selected for manual labelling), human resources can be optimized to maximize information gain with further model re-training.
- Increase user acceptance
 - The use of explainers can also help manual labelers and stakeholders to better understand the rationale behind the model results, hence help its implementation and maintenance.

Resources

- Text classification theme group final report on [ML2022 wiki](#)
 - Also recommended libraries for text classification (e.g., fasttext, transformer, top2vec)
- Collection of text classification pilot studies (from 2019-2021) on [ML wiki “Studies and Codes”](#)
- Introduction to ML-based text classification for the statistical organization in the UNECE publication [“Machine Learning for Official Statistics”](#)

Thank you!

Slido (#2909 223) for any question or comment 😊