ML 2022 Text Classification Theme Group Report

Date: November 2022

Background and objectives

Classifying textual responses into predefined categories plays an important role in the production process in the statistical organisations. Survey questionnaire and administrative registers often contain free text responses to open ended questions (e.g., job description, economy activity description). Much of information in the promising new data sources such as websites or social media also exists in the textual forms. These texts first need to be coded into international statistical classification system or internal codelist to be used and analysed for other downstream works.

Traditionally, this task used to be done manually or through a complex rule-based system, both of which are costly, time-consuming and hard to manage. With the advance of natural language processing (NLP), machine learning (ML) techniques can help statistical organisations conduct this text classification task in a more efficient way.

In the past few years, text classification has been shown to be one of most promising and popular ML application areas in the field of official statistics. For example, 9 out of 19 pilot studies under the HLG-MOS ML Project (2019-20)[1] and 5 out of 11 in pilot studies under the ONS-UNECE ML Group 2021[2] were about the text classification.

Text Classification Theme Group

With the growing maturity of expertise in this field, the ONS-UNECE ML Group 2022[3] created a sub-group primarily aimed to provide a knowledge exchange platform for those working on text classification in statistical organisations to share their works, receive feedback from peers and discuss on common challenges.

The Theme Group had a series of meeting over the year. The meetings consisted of a presentation from the sub-group members, followed by Q&A and discussion. Table below provides the overview of the meetings and presentations that took place.

Table: Overview of ML Group 2022 Text Classification Theme Group Meeting

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; naïve bayes, logistic regression, SVC, SGD, Random

[1] https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies
[2] https://statswiki.unece.org/display/ML/Machine+Learning+Group+2021
[3] https://statswiki.unece.org/display/ML/Machine+Learning+Group+2022

					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, labl2vec
7	Nov	Wrap-up meeting	-	-	-

Discussions

The presentations were given by various members on different use cases from different statistical organisations. However, there are trends, topics and concerns shared by multiple use cases as well as useful lessons learned. Below describes some of key observations made.

1. **ML-based text classification has many potential application areas in statistical organisations, even outside the production area**
    - Application of ML for classifying textual responses in survey questionnaire continues [5][4] to be a solid use case. There is clear advantage of ML-based methods over legacy methods (e.g., in terms of time, scalability).
    - ML-based methods are particularly indispensable when it comes to big data which is becoming an increasingly important data source for statistical organisations. In [1,2,6], data from web (web-scrapped data, google trend data) are used and it would be simply not feasible to manually process the data due to its size. Application on the registry data [4] is also expected to become important as more and more statistical organisations move to register-based production.
    - It is important to note that the potential of ML-based text classification is not limited to statistical production area only. For example, in [3], ML was used for triaging the

[4] The number in the square brackets refers to ID of presentations in Table 1 (e.g., [1] is the presentation "Use of ML techniques for classification problems related to CPI" from IBGE, Brazil)

enquires from clients, which improves the quality of customer service that orgnisations provide (i.e., more efficient triaging). In [6], the topic analysis based on NLP ML methods allowed statistical organisations to provide a service to policy makers and other governments (e.g., which companies are likely to be engaged in AI industry).

- It is noteworthy to mention that pre-trained language models (e.g., transformer) can be run locally to translate texts into a different language as seen in [2] which could help statistical organisations deal with multilingual microdata.

2. **NLP/ML is a fast-changing field**
    - One of difficulties for statistical organisations in this application area is that the methods and techniques in this field change fast, making it hard to stay up to date. Even few years ago, methods such as TF-IDF or Random Forest were most popular (as seen in many pilot studies). These methods, although they are still used as baseline, are increasingly challenged by more recent methods such as word embedding, transformers, and other deep learning models. Therefore, statistical organisations should constantly invest in expanding knowledge and experimenting with new methods as well as supporting staff in training of the new methods.
3. **Is it worth changing to more advanced but complex ML methods from classical ML methods that work just fine?**
    - The advent of new methods inevitably raises questions of whether one should always replace the existing (ML) methods with the latest ML methods (e.g., continue using Random Forest or replace with Deep Learning method?). Result from a quick survey among the theme group members during the kick-off meeting hinted that the advanced methods are not necessarily preferred in terms of performance (when asked “What were your experiences in using more sophisticated ML models compared to "classical" ML models?”, with only 13% responded “outperforms”). Given the investment and resources required to adapt the pipeline built around the current method, participants felt that the performance improvement of these “sophisticated” methods over “classical” methods is not enough, if not only comparable.
    - However, there are factors that could inform the decisions between advanced methods and classical methods, most notably, the complexity of the target texts. Often, job descriptions in the survey (or register) are relatively simple and concise (e.g., “cook”, “truck driver”) in which case, classical models that use word-level information might just work fine. However, as seen in [3], when the target text is relatively long and complex, more complex models seem to perform better than classical ones.
    - Also, the insights from [5] suggests that one should pay attention to the impact of how training data is built on the ML performance. For example, deterministic rule-based system relies often on vocabulary-level information (e.g., if a description contains “teacher”, then assign to a SOC code “25-000”). And if the training data are from past classification results done by this rule-based system, classical methods such as Bag-of-Word or TF-IDF which also often rely on vocabulary-level information would mimic well what rule-based system could already do. In [5], ML-based method is applied for records that could not be classified using rule-based system (so that assisted-coding can be more efficient), hence the comparison of performance should be done through those records that rule-based system fails (which tends to be longer and more complex), not those that rule-based system succeeds. Results from [1] have also shown how the use of explainers is a valuable tool that can shed light on the results predicted from ML models, even when the models are “black box”, and be used both in assisting in validating the

results obtained and pointing to sectors where the model can be fine-tuned to improve their performance.

4. **Strategies 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. The class imbalance tends to be more prominent when the target classification system has many categories (e.g., COICOP has 338 categories[5]).
    - Several strategies were observed in the presentations to address this issue. For example, in [4], oversampling (of data points from rare classes) was used, and in [5], different thresholds (threshold to decide whether to accept ML predictions or reject/send to human coders) were used for different categories to account for uncertainties associated with rare class. In [1], auxiliary data set was used to augment the data for rare classes (e.g., adding official CPI description corresponding to rare classes in the training set so that there are more descriptions that ML model can learn from).
    - 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, for example, by reviewing the data with subject matter experts and removing any classes that are not deemed important.
5. **Use of unsupervised learning models**
    - Many ML applications in statistical organisations are around the supervised learning models which rely on the labelled data. However, obtaining enough labelled data is often quite difficult, for example, when the application is for big data and the project is at the early PoC stage. The unsupervised learning models can be considered in such case as a starting point (e.g., assigning topics through unsupervised models, and use them as labels after some validation as in [6]).
6. **Beyond Proof-of-Concept**
    - There have been enormous progresses made in the field of NLP/ML in the past few years, making things that could not been imagined to be done by machines possible. However, there are many challenges to turn a potential solution into a real solution in production[6].
    - One of important considerations is the maintenance strategy which is crucial for the sustainability of the model. Preparing in advance the monitoring and re-training plan could increase the acceptance by stakeholders and facilitate the transition from PoC stage to production stage. Hence, it is recommended to start considering the retraining plans while developing the solution. For example, in [4], 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). In this way, human resources can be optimized to maximize information gain with further model re-training. The use of explainers as in [1] can also help manual labelers and stakeholders to better understand the rational behind the model results, hence help its implementation and maintenance.

[5] At the sub-classes level in 2018 version (https://unstats.un.org/unsd/classifications/unsdclassifications/COICOP_2018_-_pre-edited_white_cover_version_-_2018-12-26.pdf)

[6] For more, see Chapter 5 "Journey from Machine Learning Experiment to Production" from Machine Learning for Official Statistics (https://unece.org/sites/default/files/2022-01/ECECESSTAT20216.pdf)

Resources

Some of resources and key libraries used in the works in the Table are listed below:

- LIME: Local Interpretable Model-agnostic Explanations (used in [1])
- fastText: A library for efficient learning of word representations and sentence classification (used in [2]). It provides an easy and efficient way to apply, among other, the following methods:
    - Pre-trained word vectors: English word vectors · fastText and for 157 languages
    - Python resources: (and command line)
        - Text classification · fastText
        - Word representations · fastText
    - R resources: fastRtext package wrapper
        - Text classification (Supervised learning)
        - Word representations (Unsupervised learning)
- Multilingual models: overall Hugging face has more than 90K models and you can choose the most appropriate based on the task at hand (BERT is one of the most downloaded ones). Most of them also come with snippets so it's easy to get started
    - The Language Technology Research Group at the University of Helsinki has more than 1400+ machine translation models covering a variety of languages
    - Another one is Google T5 model. But it only covers a small set of languages
- Unsupervised topic modelling and text classification (used in [6])
    - Top2Vec: topic modelling and semantic search
    - Lab2Vec: unsupervised classification