SPEAKING THE SAME LANGUAGE:
A machine learning approach to classify Burning Glass skills

Julie Lassébie (OECD)

Joint with Luca Marcolin (OECD), Marieke Vandeweyer (OECD), Benjamin Vignal (ENSAE)

May 2021

Rationale and overview

- Burning Glass data contains information retrieved from (millions of) online job vacancies (US, UK, CAN, SGP, AUS, NZL)
- For each job posting, it lists skill requirements, among other useful variables
- Very rich information that can be used to study differences in skill demand over time and across sectors, wage returns to different skills, …
  - Notably in the context of the Great Recession (e.g. Hershbein and Kahn, 2018)

- It contains **more than 17,000 distinct "skills"** keywords across countries and years
  - Empirical analyses require a (much) lower number of categories
  - Several synonyms or close concepts should be grouped together
- We want to **classify keywords into a smaller number of categories, based on their meaning**

Working with skill information from Burning Glass: insights from the literature

- **Indirect measures** of skill requirements: education and experience (Blair and Deming, 2020)
- A number of studies use a **restricted number of categories** to identify certain skills (Hershbein and Kahn, 2018; Deming and Kahn, 2018; Deming and Noray, 2020)
  - Number of categories ranges from 2 to 10
  - For each category, the authors list the different accepted keywords
  - To infer demand for one category, they consider the number of job ads listing at least one of the keywords
  - The researcher needs to specify ex-ante which categories are/will be important in the labour market and determine the list of accepted keywords

Our approach

- We want to **classify *all* skill keywords into an existing taxonomy, based on the meaning** of BG skills and the different categories
  - Does not need to specify ex-ante important categories
  - Uses the **existing knowledge on skills concepts**
  - Refers to a **framework validated and understood by** labour market and education **experts**, statistical agencies, and stable over time.

- We use a **supervised machine learning** approach

Pre-requisites

- Construction of the **target taxonomy**
  - Starting point: **O*NET taxonomy**
  - Some changes: new **digital skills categories,** merged categories that were too close in meaning
  - Our O*NET+ final taxonomy contains **61 categories**
  - These categories can be grouped ex-post for empirical analyses, when necessary
- **Definitions** of skill keywords
  - Retrieved from ESCO when available
  - Scrapped from Wikipedia otherwise

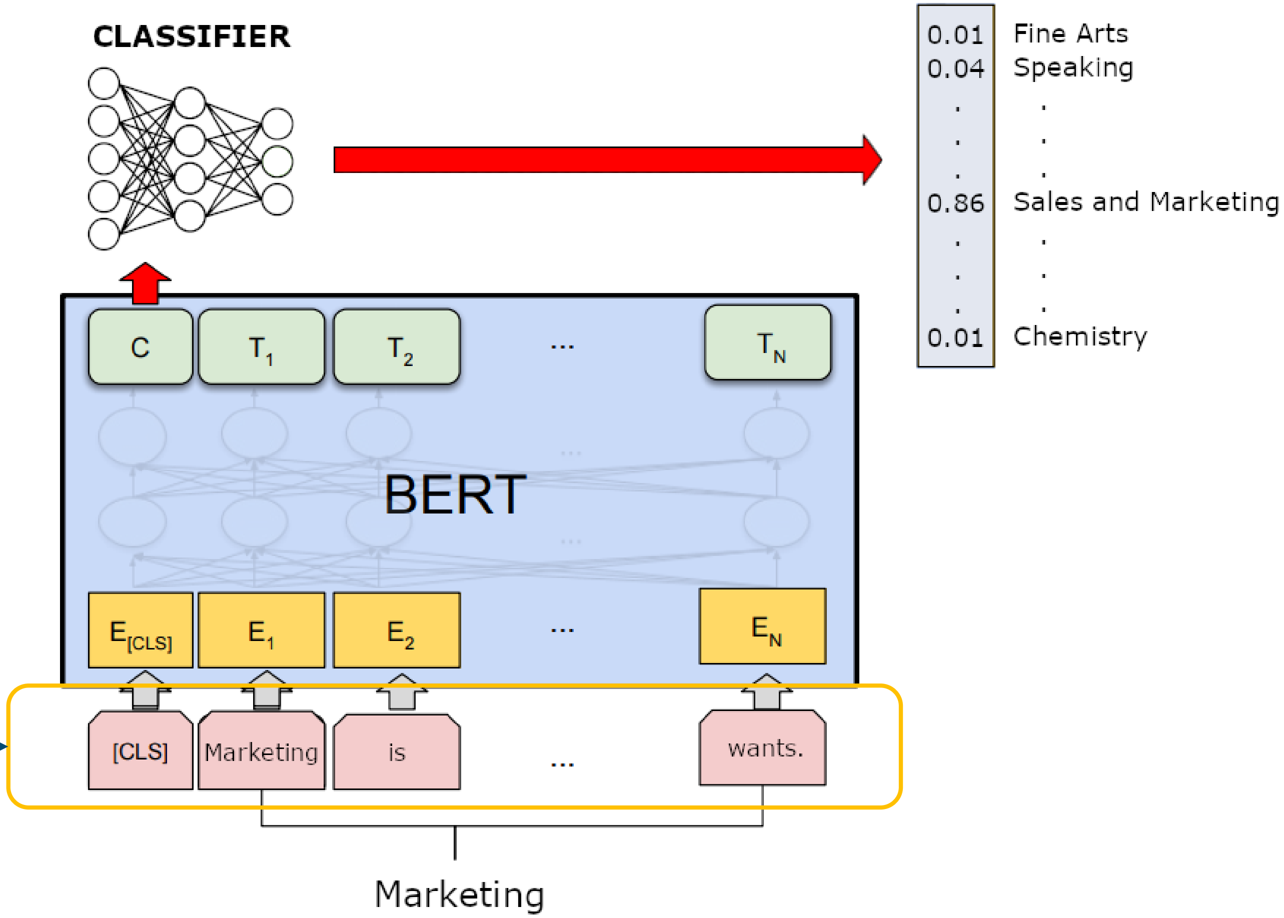
BERT model

- To classify a BG skill we rely on **BERT** (Bidirectional Encoder Representations from Transformers), developed in 2018 by Google AI.
  - Particularly suited for sentence classification (spam/non-spam, sentiment analysis, etc.)

- BERT in a nutshell:
  - **Pre-trained** on a large corpus: readily available word-embeddings (= vector representation of words).
  - Word embeddings are then refined to take context into account and tailored to the classification task
  - *Step 1:* Takes **definitions of skills** and transforms them into **word embeddings**
  - *Step 2:* **Classification of each vector** (associated to each skill) into one of the 61 categories
    - Applying a softmax classifier (akin to multinomial logistic regression)
    - Using a labelled dataset (or training set)

Example

- Marketing : “Marketing is the study and management of exchange relationships. It is the business process of identifying, anticipating and satisfying customers' needs and wants. […]”

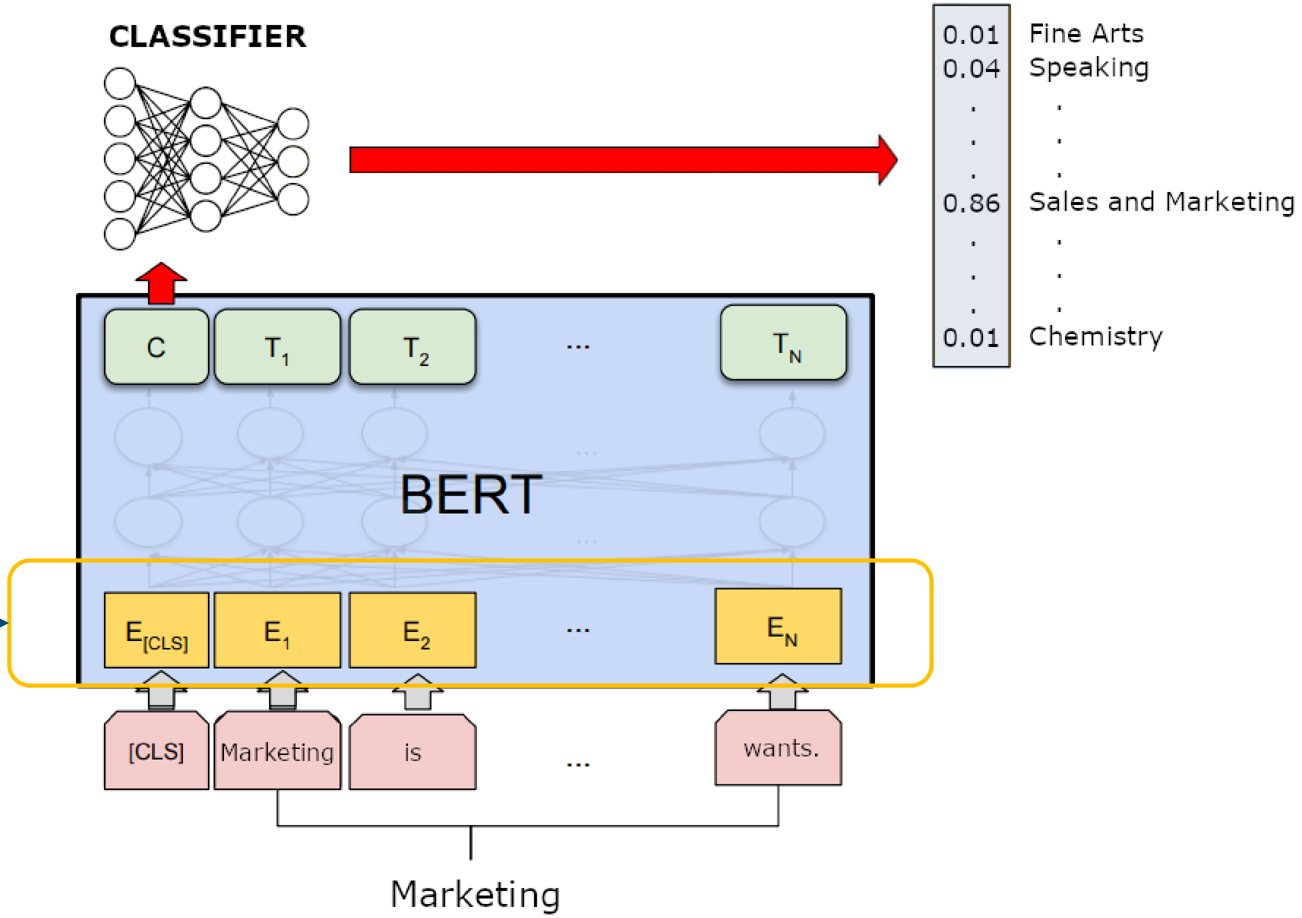


0.01	Fine Arts
0.04	Speaking
.	.
.	.
.	.
0.86	Sales and Marketing
.	.
.	.
.	.
0.01	Chemistry

Textual definition of “Marketing”

Example

- Marketing : “Marketing is the study and management of exchange relationships. It is the business process of identifying, anticipating and satisfying customers' needs and wants. […]”

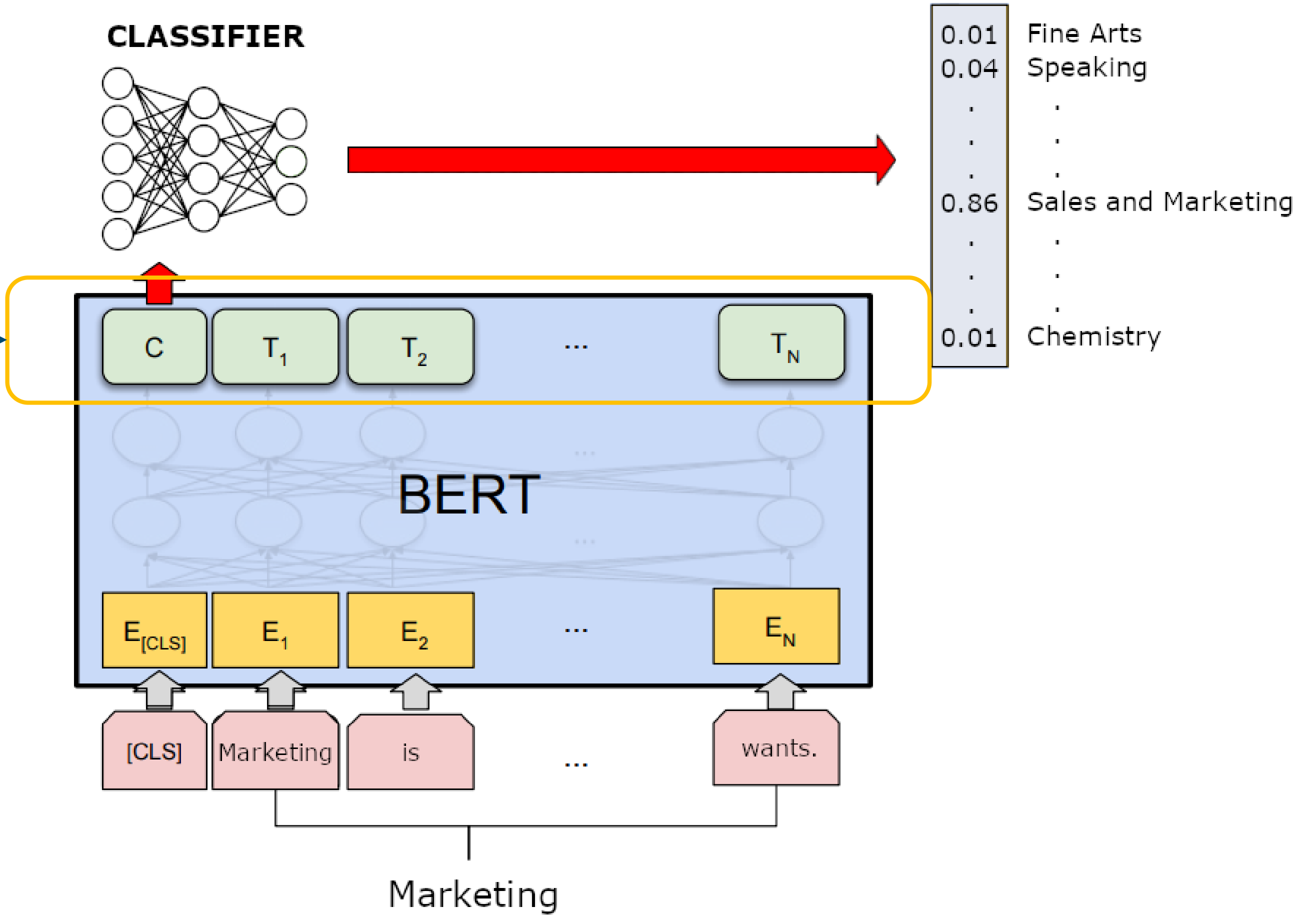


Individual embedding of each word

Example

- Marketing : “Marketing is the study and management of exchange relationships. It is the business process of identifying, anticipating and satisfying customers' needs and wants. […]”

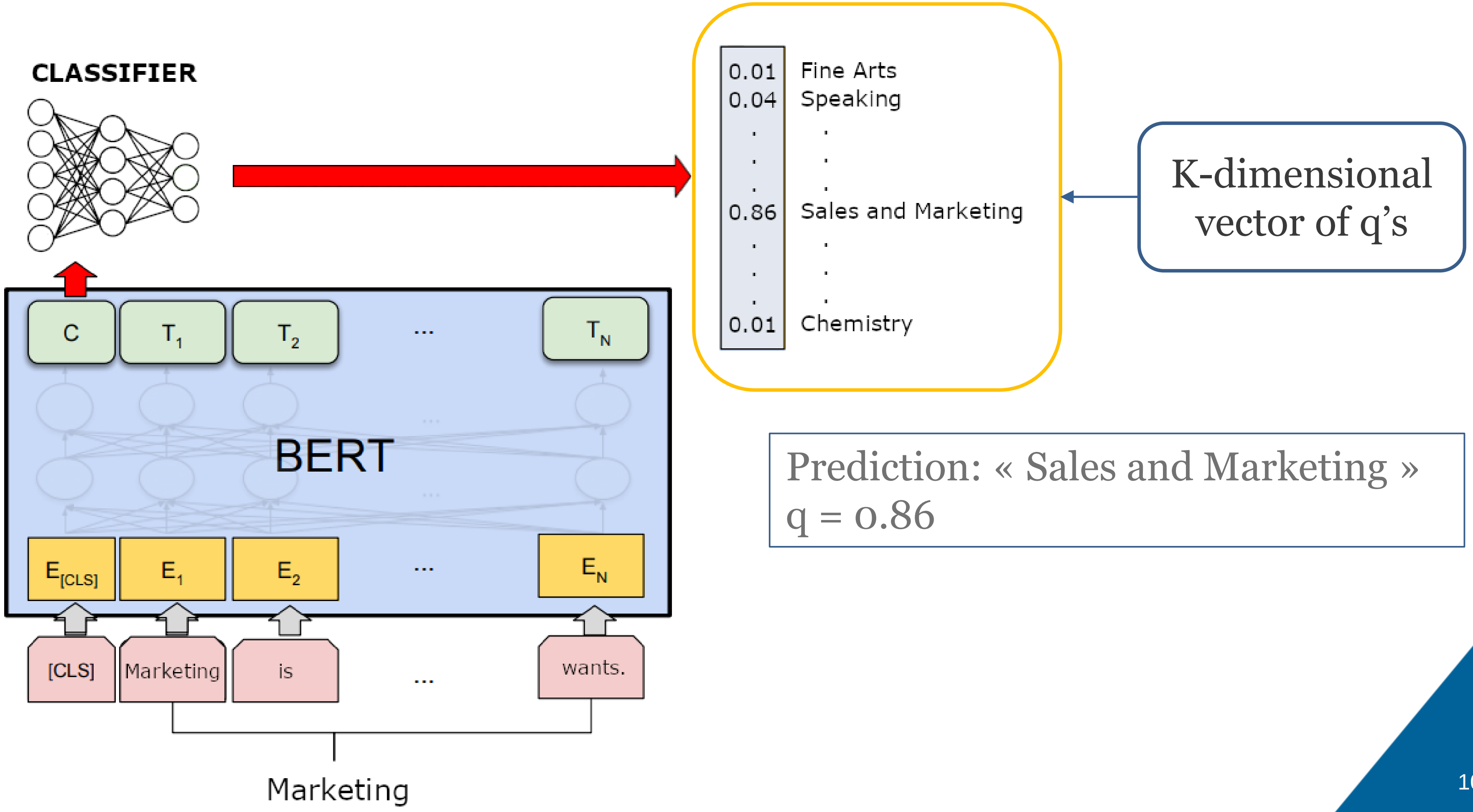
Contextualized embedding of each word



0.01	Fine Arts
0.04	Speaking
.	.
.	.
.	.
0.86	Sales and Marketing
.	.
.	.
.	.
0.01	Chemistry

Example

- Marketing : “Marketing is the study and management of exchange relationships. It is the business process of identifying, anticipating and satisfying customers' needs and wants. […]”

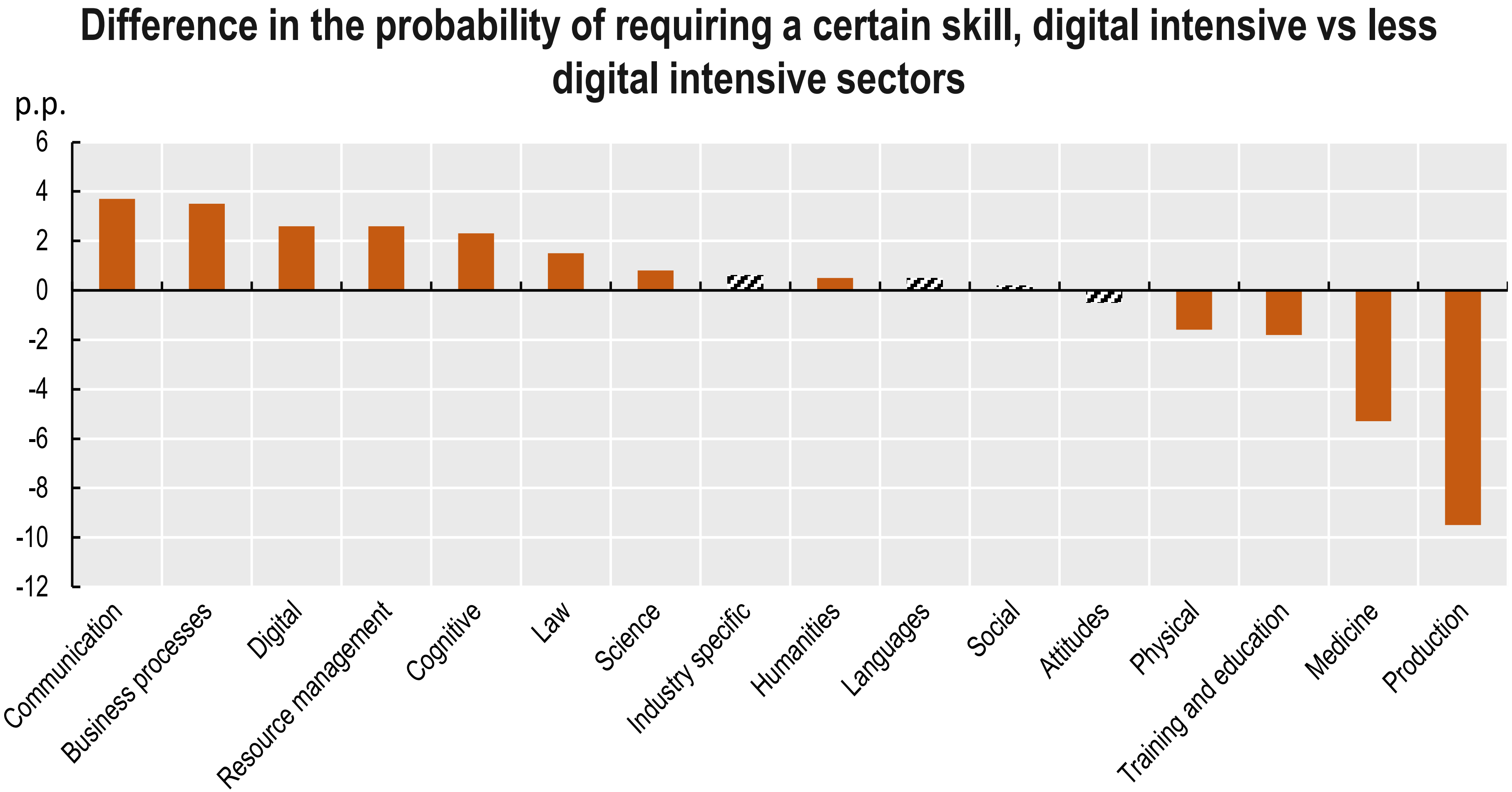


Prediction: « Sales and Marketing »
q = 0.86

Discussion of the results and internal validity

- 17 331 skills grouped into 61 categories
  - 180 keywords not classified when definition not available
  - Most populated categories: “Medicine and Dentistry” (3 895 skills), “Management of Financial Resources” (1070 skills), “Biology” (1 044 skills)
- **Accuracy**: frequency of model *correctly* allocating a skill to a category ⇒ check results of the algorithm against a **test set:**
  - 150 random skills, manually classified into categories. Compare this with the allocation produced by the algorithm.
- The accuracy measured on the test set is comprised between **74% and 90%**
  - In general, the BERT model applied to different problems achieved accuracies between 65 - 95% (mostly tested for binary classification problems that are much simpler to deal with).
  - Many BG skills are vague or incongruous (e.g. Bowling, Human Guides, VTPSUHM7, HORVIP, etc.)
  - Definitions of keywords may not be sufficiently long or precise for the full identification of the skill

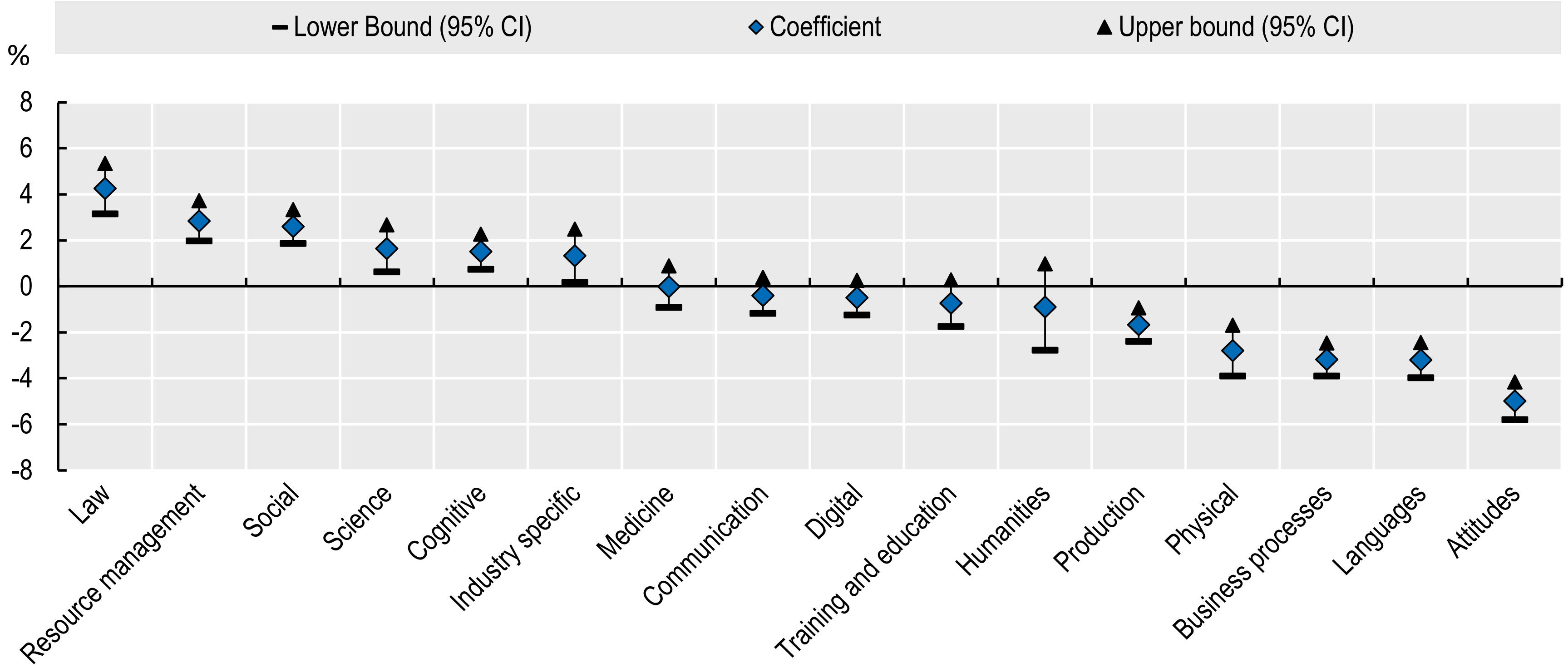
Assessing (indirectly) external validity I



Note: The graph plots the marginal effect of an OLS regression, where the probability that a job advertisement requires at least one skill in a given skill category is regressed on an indicator variable with value 1 if the job is advertised in a digital intensive sectors according to Calvino et al. (2018[54]) and zero otherwise, and dummies for State, 2-digit U.S. SOC 2010 occupation and employer name. Shaded bars identify differences (coefficients) which are not statistically significant at the 5% confidence level based on robust standard errors.
Source: OECD estimations based on Burning Glass Technologies data for CAN (2018).

Assessing (indirectly) external validity II

Percentage change in hourly wages if the job posting requires at least one skill in the category, everything else held constant.



Note: The graph plots the coefficients of a single estimation on Canadian 2018 job postings data, where the logarithm of posted hourly wages is regressed on required experience, dummies for educational attainment, dummies for all skill requirements (one for each of the 16 "Broad" categories presented above), State, 2-digit ISIC rev.4 industry and 2-digit U.S. SOC 2010 occupational dummies. Dummies for skill requirements take value 1 if the job posting requires at least one skill falling into the category on the x-axis.
Source: OECD estimations based on Burning Glass Technologies data for CAN (2018).

Conclusion

- We needed to **classify skill keywords** appearing in BG job ads **into a smaller number of fixed categories, based on the meaning of keywords and categories**
- We use **BERT**, a NLP algorithm particularly suited for sentence classification
- The results of the classification show a **satisfactory accuracy and strong correlation with other trusted sources of information**
- Next steps: using this work to **analyse changes in skill demand**

Thank you

Julie.Lassebie@oecd.org

ANNEX – TECHNICAL DETAILS

Skills in Burning Glass data

- **17 348 different skills** across all years and countries, including **8 595 that are common to all 6 countries**
- Most job ads (~98%) contain 1 to 20 skills

	AUS-NZL	CAN	SGP	UK	US
Job ads	6 974 051	6 487 666	3 489 531	52 393 082	148 117 657
Unique skills	11 608	12 979	10 881	12 238	15 847
Average number of skills per ad	5.4	9.4	7.4	5.7	8.6

Source: OECD calculations on Burning Glass Data

Pre-requisites II

- 3 categories filled manually:
  - Industry Specific Knowledge: for any keyword of the form “xxx Industry Knowledge” (297 keywords)
  - Local Language: country-specific, for now English
  - Foreign Languages: any identified language other than Local Language
- Definitions
  - From ESCO when possible (674 keywords)
  - From Wikipedia entry (for all other keywords)

BERT

- BERT uses Transformer, a bidirectional training algorithm using Masked Language Modelling
  - Research shows that Transformers tends to be superior in quality while being more parallelizable and requiring significantly less time to train
- Pre-trained on massive amount of (unlabeled) data (Wikipedia inter alia): readily available (contextualized) word-embeddings
  - 12 layers of encoders
- Supervised Learning step: we run the BERT algorithm (again) to improve word-embedding for our classification task (using labelled dataset)
  - Takes sentences (definitions) as inputs
  - Produces word-embeddings for each word in the sentence + an extra vector as output (Classification token)
  - Classification tasks using BERT are performed by adding a classification layer on top of the Transformer output to fine-tune the Classification token.
  - Classification performed using a traditional softmax classifier (akin to multinomial logistic regression)

How the classification works

- The output for the multinomial logistic regression (also called **softmax classifier**) is a K-components vector that sums to **1,** with K= number of categories in the ONET+ taxonomy.
  - Note : some skills (languages, « industry knowledge ») can be automatically matched via dedicated exhaustive lists, so effectively K= **58** categories.
- These values can be seen as the “probability” for the skill to belong to the corresponding category.
- The maximum value, which we call **q**, defines the predicted category. **q** is the **confidence** the model has in its prediction.

Training the model

- Two objectives in the construction of the training set:
  - **Homogeneity** (approximately same number of skills in each category)
  - **Universality** (representative content in each category)
- Our training set is composed of:
  - the 200 most frequent skills in BG
  - 100 randomly sampled skills
  - 184 out-of-sample skills (specifically chosen to improve homogeneity and universality for some categories)