gomez_2022_large_scale_analysis_of_open_mooc_reviews_to_support_learners_course_selection

Year

2022

Author(s)

Manuel J. Gomez and Mario Calderon and Victor Sanchez and Felix J. Garcia Clemente and Jose A. Ruiperez-Valiente

Title

Large scale analysis of open MOOC reviews to support learners' course selection

Venue

Expert Systems with Applications

Topic labeling

Fully automated & Partially automated

Focus

Secondary

Type of contribution

Established approaches

Underlying technique

Top-words selection and manual category assignment

Topic labeling parameters

\

Label generation

Approach 1 - Qualitative description model

Each topic is labeled by its three most important keywords (e.g., ''simple_easy_effective''). As we can observe, the most frequent topics are ''informative_easy_fun'' (12.2%), ''basic_beginner_introduction'' (8.8%) and ''personal_authentic_productive'' (8.3%), and the less frequent topics are ''real_worth_life'' (5.3%) and ''slide_powerpoint_visual'' (5.0%).

Approach 2 - Content model

In this case, we have assigned a label to each topic based on the existing content categories in the corpus and the keywords of each topic.

A summary of each topic, including its name, description, and the five most important words related, can be found in Table 2.

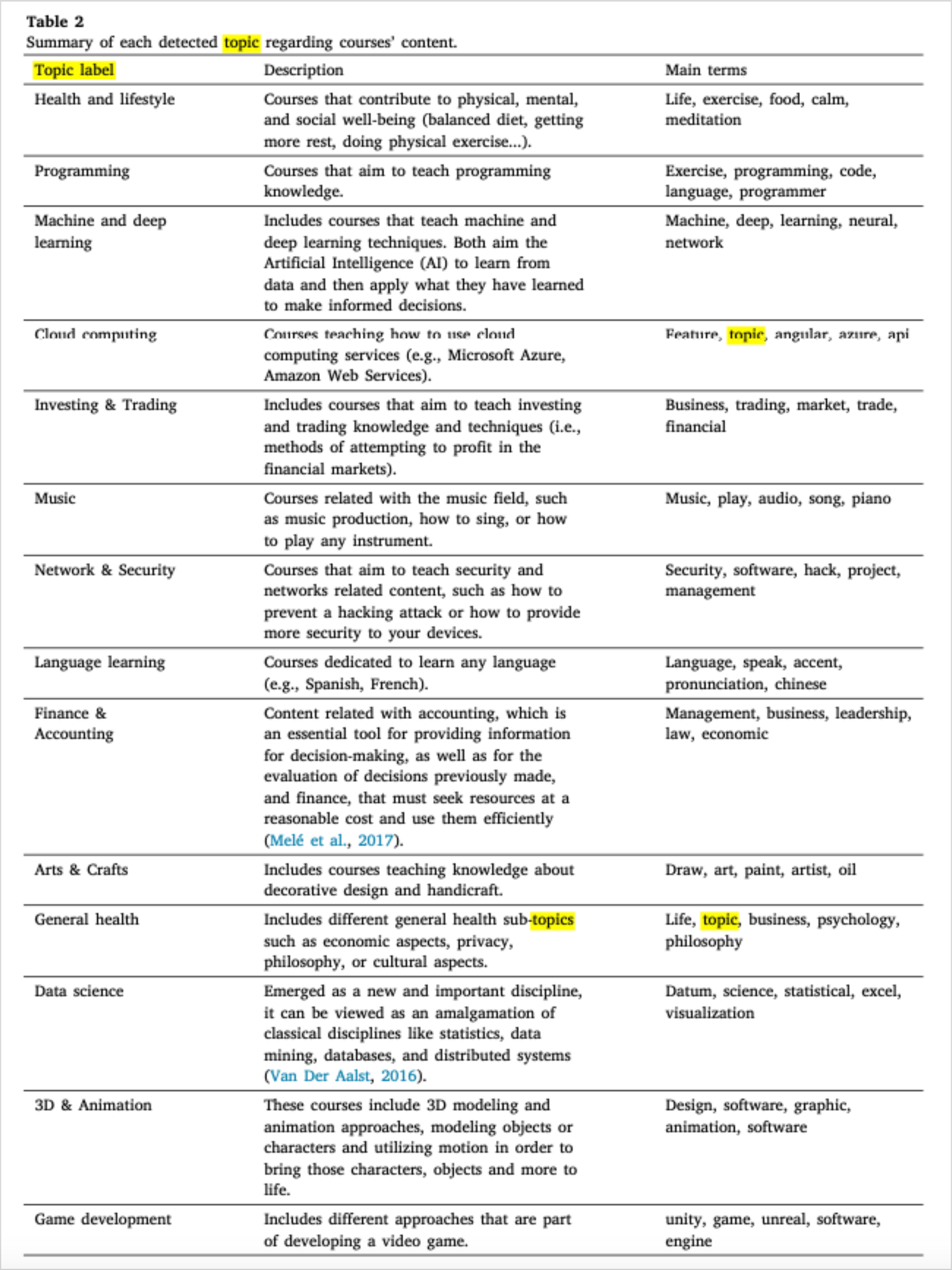

Table 2
Summary of each detected topic regarding courses' content.

Topic label	Description	Main terms
Health and lifestyle	Courses that contribute to physical, mental, and social well-being (balanced diet, getting more rest, doing physical exercise...).	Life, exercise, food, calm, meditation
Programming	Courses that aim to teach programming knowledge.	Exercise, programming, code, language, programmer
Machine and deep learning	Includes courses that teach machine and deep learning techniques. Both aim the Artificial Intelligence (AI) to learn from data and then apply what they have learned to make informed decisions.	Machine, deep, learning, neural, network
Cloud computing	Courses teaching how to use cloud computing services (e.g., Microsoft Azure, Amazon Web Services).	Feature, topic, angular, azure, api
Investing & Trading	Includes courses that aim to teach investing and trading knowledge and techniques (i.e., methods of attempting to profit in the financial markets).	Business, trading, market, trade, financial
Music	Courses related with the music field, such as music production, how to sing, or how to play any instrument.	Music, play, audio, song, piano
Network & Security	Courses that aim to teach security and networks related content, such as how to prevent a hacking attack or how to provide more security to your devices.	Security, software, hack, project, management
Language learning	Courses dedicated to learn any language (e.g., Spanish, French).	Language, speak, accent, pronunciation, chinese
Finance & Accounting	Content related with accounting, which is an essential tool for providing information for decision-making, as well as for the evaluation of decisions previously made, and finance, that must seek resources at a reasonable cost and use them efficiently (Melé et al., 2017).	Management, business, leadership, law, economic
Arts & Crafts	Includes courses teaching knowledge about decorative design and handicraft.	Draw, art, paint, artist, oil
General health	Includes different general health sub-topics such as economic aspects, privacy, philosophy, or cultural aspects.	Life, topic, business, psychology, philosophy
Data science	Emerged as a new and important discipline, it can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems (Van Der Aalst, 2016).	Datum, science, statistical, excel, visualization
3D & Animation	These courses include 3D modeling and animation approaches, modeling objects or characters and utilizing motion in order to bring those characters, objects and more to life.	Design, software, graphic, animation, software
Game development	Includes different approaches that are part of developing a video game.	unity, game, unreal, software, engine

Motivation

\

Topic modeling

LDA (two models, one for QualitativeDescription and one for Content, see `pre-processing`)

Topic modeling parameters

Nr of topics (k): 2 to 25

Nr. of topics

28 (14 per model)

Label

1. Top three topic words linked by underscores
2. Single or multi word topic content categories extracted from the dataset

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Recommender systems (for MOOCs)

Dataset: MOOCs courses & reviews

Problem statement

We believe that there is an opportunity to leverage available Massive Open Online Course

(MOOC) reviews in order to build simpler and more transparent reviewing systems, allowing users to really identify the best courses out there.
Specifically, in our research we analyze 2.4 million reviews from five different platforms in order to determine the following:

1. if the numeric ratings provide discriminant information to learners
2. if NLP-driven sentiment analysis on textual reviews could provide valuable information to learners
3. if we can leverage NLP-driven topic finding techniques to infer themes that could be important for learners
4. if we can use these models to effectively characterize MOOCs based on the open reviews.

Corpus

Origin: Various MOOCs providers
Nr. of documents: 2.411.440
Details:

- Course and review data from Udemy, Coursera, Domestika, Platzi, Crehana

Document

- For each review: URL, review, rating of the review (from one to five), the platform, the username, the date of the review, and the identifier of the related course.
- For each course: the course identifier, URL, title, platform related with the course, the content category of the course, and the name of the teacher.

Pre-processing

- Removal of commas, special characters, unnecessary URLs, numbers, or additional space characters
- Removal of stop words
- Lemmatization
- Since we needed to identify the language of each review, and this information is not available in the original metadata, we identified it using the raw text of each review using the Fasttext library.
- The most frequent words are collected (words that appear more than 500 times in the entire collection), and we manually classified every word within two different categories:
    - QualitativeDescription: if the word is related to a qualitative description of the course (e.g., easy, clear, practical). This category includes 357 different words.

- Content: if the word is related to the content of the course itself (e.g., machine, yoga, cooking). This category includes 759 different words.
- Every word not matching any of those two categories will be excluded from our later analysis.

```
@article{gomez_2022_large_scale_analysis_of_open_mooc_reviews_to_support_learne
rs_course_selection,
  abstract = {The recent pandemic has changed the way we see education. During
recent years, Massive Open Online Course (MOOC) providers, such as Coursera or
edX, are reporting millions of new users signing up on their platforms. Though
online review systems are standard among many verticals, no standardized or
fully decentralized review systems exist in the MOOC ecosystem. In this vein,
we believe that there is an opportunity to leverage available open MOOC reviews
in order to build simpler and more transparent reviewing systems, allowing
users to really identify the best courses out there. Specifically, in our
research we analyze 2.4 million reviews (which is the largest MOOC reviews
dataset used until now) from five different platforms in order to determine the
following: (1) if the numeric ratings provide discriminant information to
learners, (2) if NLP-driven sentiment analysis on textual reviews could provide
valuable information to learners, (3) if we can leverage NLP-driven topic
finding techniques to infer themes that could be important for learners, and
(4) if we can use these models to effectively characterize MOOCs based on the
open reviews. Results show that numeric ratings are clearly biased (63% of them
are 5-star ratings), and the topic modeling reveals some interesting topics
related with course advertisements, the real applicability, or the difficulty
of the different courses.},
  author = {Manuel J. Gomez and Mario Calder{\'o}n and Victor S{\'a}nchez and
F{\'e}lix J. Garc{\'\i}a Clemente and Jos{\'e} A. Ruip{\'e}rez-Valiente},
  date-added = {2023-03-19 19:52:16 +0100},
  date-modified = {2023-03-19 19:52:16 +0100},
  doi = {https://doi.org/10.1016/j.eswa.2022.118400},
  issn = {0957-4174},
  journal = {Expert Systems with Applications},
  keywords = {Massive Open Online Courses, Natural language processing,
Sentiment analysis, Recommendation systems, Online education},
  pages = {118400},
  title = {Large scale analysis of open MOOC reviews to support learners'
course selection},
  url = {https://www.sciencedirect.com/science/article/pii/S0957417422015081},
  volume = {210},
  year = {2022}}
```

#Thesis/Papers/Initial