2023/2024

Data Science Project Report

AI TEXTUAL FEEDBACK ANALYSIS: WHAT DO THE ALUMNI THINK OF POLYTECH'S COURSES?

PRESENTED BY

Alexandre Deloire - Rémi Jorge

Jiayi He - Charlène Morchipont

Context

The highly selective engineering school of Polytech Montpellier conducts surveys among its alumni to gather information regarding their studies and professional integration. A study at 6 months, 18 months, and 30 months after graduation is carried out by the school.

In addition to single or multiple-choice questions, textual feedback is also provided by Polytech Montpellier graduates during these surveys. This data has been relatively understudied until now.

We have chosen to build revolutionary AI to extract the textual responses, organize them, and analyze them.

What are we going to study?

We will study the textual feedback from former students of Polytech Montpellier regarding their education and professional integration, collected during surveys conducted at 6 months, 18 months, and 30 months after obtaining their diploma.

What problem are we going to solve?

Many of the textual responses address feedback on the education, particularly the courses. Former students express their opinions on courses they find useful, useless, missing, and those that could have been beneficial or need strengthening for better professional integration. We propose to study these responses to determine, for each field of study, which courses were deemed useful, useless, etc.

Goals

- Extract the data from the surveys
- Clean and organize the textual data collected from graduates to eliminate errors and inconsistencies.
- Conduct an overall analysis of the data to identify suitable analysis methods.
- Implement the analysis methods.
- Conduct a detailed analysis of textual data and extract results.
- Propose an interactive interface for decision-makers.

Introduction

In the hallways of Polytech, one often hears students complaining about certain courses, while others would have preferred a reinforcement of certain subjects. Indeed, courses are a major concern for students, as the primary reason for attending Polytech is, obviously, the education provided. Analyzing this data is therefore a crucial asset to better understand the students' opinions, enabling decisions to be made regarding specific courses for each field of study and, consequently, improving the education at Polytech.

Initially, it is imperative to have an overview of the data we have because we know nothing about this information. Here is how we proceeded for this step:

Let's talk about the data:

The data was downloaded from the internet. Here are some characteristics about it, and some pose immediate issues:

- The data is stored in separate files.
- The data is in Excel format but from different years.
- The file encoding is not the same; some are in Latin-1, others in utf-8.

We developed a Python program to extract, for each file, a list of these questions along with the characteristics of each question, including:

- The question
- The type of response
- General information about the responses, such as if it is a category or binary: enumeration of possible values, if it is a number, the min and max, etc.
- The number of responses
- The number of missing responses
- The number of unique values

With the extracted information, we were able to discuss the best questions to use and the methods we would employ for this textual study. The question extraction program is provided in the appendix.

These questions caught our attention:

- Which courses do you find most useful for your profession and professional integration?
- Among the courses provided by the school, which ones deserve to be deepened or reinforced?
- Which courses, absent from your education, would have been useful to you?
- Which courses, present in your education, seem unnecessary?

Before applying the analysis methods presented in this report, it was necessary to clean the data, which was almost unusable. In the following part, we will discuss the problems encountered and the solutions put forward.

Extract, Transform, Load (ETL)

The Excel formats from different years pose a problem because the file encoding is not the same. Worse than that, the 2020 file includes encoding errors that make sentences with accents almost irrecoverable. Here is an example of a sentence that can be found:

“la comptabilitÈ matiËre, la comptabilitÈ en gÈnÈral, du droit juridique”

There is an encoding issue, with acute accents becoming uppercase and grave accents.

Another problem is that the files have passed through several operating systems, including Windows, macOS, and Linux, and there are character escape sequence "\r" and "\n" reading errors, which, in addition to the encoding error, result in the appearance of "_x000D_".

Furthermore, between each file, the wording of the questions changes slightly, sometimes with the addition or deletion of a word, and new questions may appear while others disappear. Therefore, it is necessary to standardize the questions so that during the analysis, the same questions can be identified across different files.

Next, it is necessary to convert all the files into a consistent and recent format.

To achieve this, we created a Python script named "clean_up_excel.py," which handles this cleaning and conversion. Once the cleaning is done, it is essential to have a simple way to extract the data and columns we want to process in a CSV format, facilitating analysis and training. For this purpose, we created a Python script named "xls_to_csv.py," configurable with column names, which extracts all the necessary data into a new file.

With this completed, we can now begin to exploit the data for analysis.

Data Overview

Before any in-depth analysis, it is important to obtain some basic information about our data. We created simple graphs to precisely understand the nature of the responses. Decision-makers have exactly the same questions.

For example:

- How many responses are there per year, per field of study, per gender?
- How many of them answered the textual questions? etc...

By analyzing the graphs, we observe that each field of study and gender is well represented each year, indicating that we can draw meaningful conclusions from our in-depth analyses below.

We also created some graphs for non-textual questions to later correlate them with the results of textual questions.

The graphs are provided with our interactive decision-making tool and in the appendix.

Word Clouds

The analysis of textual data is a complex task; it involves not numbers but rather an accumulation of words that together create meaning. The goal is to represent the data in the simplest, most intuitive way so that the user of our application can have a reliable overview of their survey in our case.

Therefore, we decided that word clouds would be an excellent way to represent student's opinions on the courses of their education. Our word clouds pertain to two distinct and highly relevant questions for analysis:

- Which courses do you find the most useful for your profession and professional integration?
- Among the provided courses, which ones deserve to be strengthened?

The courses delivered at Polytech are diverse. Some courses are common to several fields of study but differ slightly in terms of content. Several challenges had to be overcome to achieve good results.

The construction of the word clouds was carried out in several stages:

- Prioritize word coherence by gathering information in advance about the courses of different fields of study. It was imperative to gather information about the subjects taught in other fields of study first. Identify courses name composed of multiple words ("mécanique des fluides," "chimie organique," etc.). Find possible abbreviations to match the same words (e.g., "rdd" -> "resistance des matériaux").
- Create lists of words, grouping courses from the same UE (Unité d'Enseignement) or theme. This allows counting words that mean the same thing together and makes their weight more reliable, also providing better visibility.
- Refinement phase: Move from substantial groupings to smaller ones that genuinely represent useful sub-ideas rather than a new visualization of the courses brochures.
- The final step was to retrieve, for each field of study, the top 5 most mentioned courses.

Tools used:

- Regular expressions to identify subjects composed of multiple words.
- NLTK library, stopword, to ignore conjunctions in a text like "et" (and), as these are often repeated words, keeping them would have caused issues.

Results

The word cloud representing useful courses for professional integration is very interesting. We can genuinely see subjects that stand out, such as, for example, in IG (Computer science and Management), web, and software, which are indeed key courses of the program. No inconsistency in responses is visible, and we have a good understanding for each field of study of the most important courses.

As for the word cloud representing courses that deserve to be deepened, we notice that the main subjects also appear and are often worthy of further exploration.

For more precise and detailed results, it is interesting to look at the outcomes of the k-means method that we also employed.

Analysis of the satisfaction

Once the word clouds were created, we wanted to understand the general trends reflected in the written comments. The question we wanted to study is:

"Your feedbacks and comments regarding your professional integration."

Indeed, it is interesting to know whether the feedback is generally positive or negative because professional integration is linked to the courses. These courses guide us more or less in specific domains, and the knowledge acquired or not greatly influences success in the professional world. With this data, we can easily identify if there is an issue in a field of study. Perhaps the courses are no longer meeting the expectations of the working world, etc.

Challenges encountered

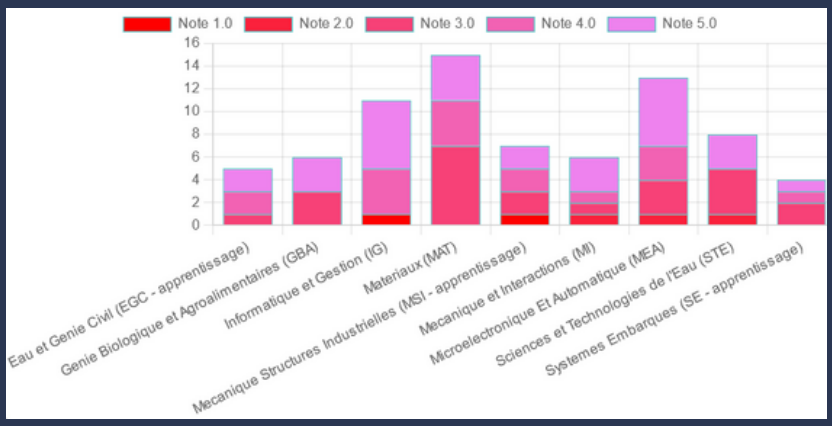
For this analysis, we didn't have a knowledge base as we did for the word clouds because the comments are highly random. Therefore, we had to find a way to process this large amount of data efficiently and make sense of the comments.

The solution we found to achieve this is the use of pre-trained neural networks for sentiment analysis. There are several available, but not all of them are reliable, and finding a precise and consistent tool was not straightforward.

Initially, we considered using Textblob and NLTK. Unfortunately, the reliability test results were inconclusive. A highly praising comment was classified as neutral, as was a very negative comment. So, we had to research and test other possible methods before finding the right tool.

In the end, we used the transformers module with a pipeline, which allowed us to classify comments from 1 star (negative) to 5 stars (positive). The results were very reliable, and we were able to create a graph representing satisfaction from a professional integration perspective.

Result



Graph representing students' satisfaction regarding professional integration based on written comments.

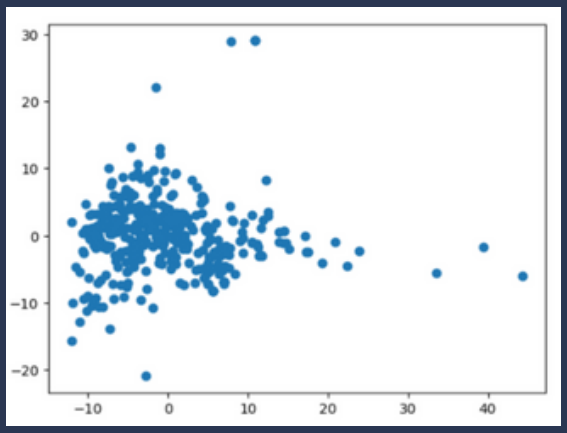
The results obtained are very good because we have a means of quantifying satisfaction effectively. Thanks to this, we can focus on the analysis of a specific field of study that may have received more negative feedback and diagnose the problem using the word clouds described earlier. For 2022, for example, we observe that 50% of the responses from the MAT (Materials Science) field are rated 3 stars, so it would be interesting to analyze this field precisely to understand the reasons behind this rating.

Classification

How to approach the analysis of a large quantity of textual data? The goal is to group comments to identify similarities and obtain an overall view of the expressed opinions. Therefore, we are faced with a classification problem.

Converting text to vectors

After data selection and cleaning, the first step is to transform the text into a format understandable by a computer. To do this, we convert the text into a vector, also called embedding. We use the neural network trained by Spacy for French, opting for its large version. After comparing several neural networks, the large version of Spacy proved to be effective for this step, which motivated our choice. The result is a vector in 300 dimensions.

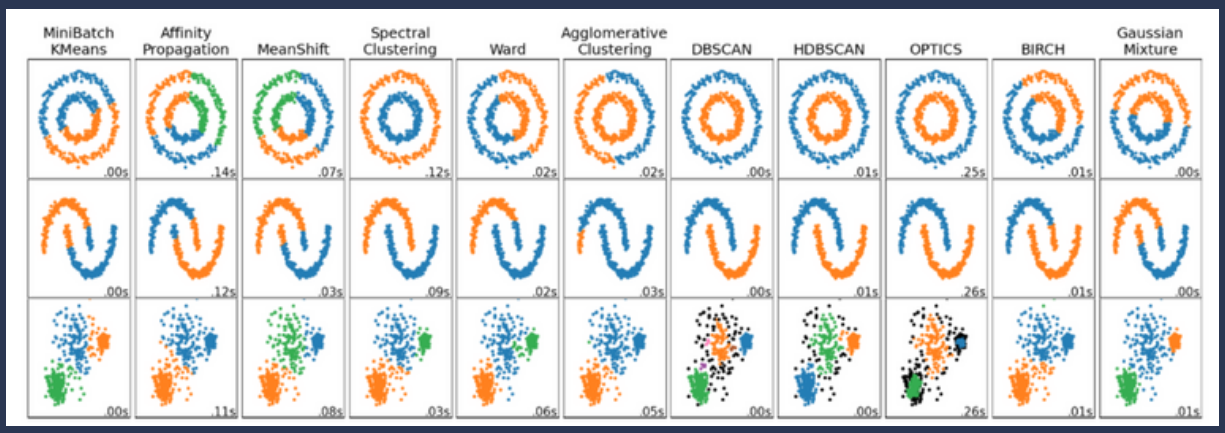


In order to obtain a visual representation, we use Principal Component Analysis (PCA) to reduce the dimension to 2. In the example below, focusing on the question: "Which courses do you find most useful for your profession and professional integration?", we have selected the STE sample.

Visualization of the vectors projected in 2 dimensions.

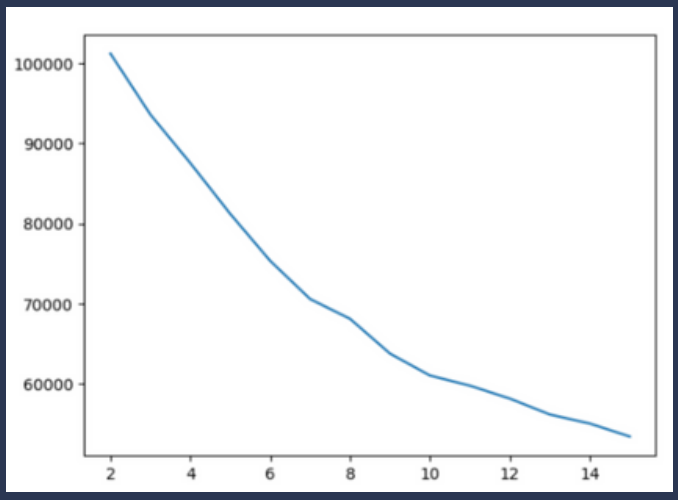
Choosing a model and the parameters

The next step is to choose the classification model. Several algorithms exist for unsupervised classification. Although we were already familiar with the K-means algorithm, we wanted to compare its performance with other methods. After various trials with diverse data, the results indicate that K-means provides among the best classification performance, justifying our choice.



Different methods of unsupervised classification and their performance based on the type of data.

The K-means method requires predefining the number of clusters to form. This is a crucial parameter, as an inappropriate choice can lead to poor results. We explored several methods to calculate the optimal number of clusters. The first method involves examining the average total inertia as a function of the number of clusters. Since K-means aims to minimize inertia, this proves to be a good indicator. It is then a matter of identifying an "elbow" to determine the optimal number.

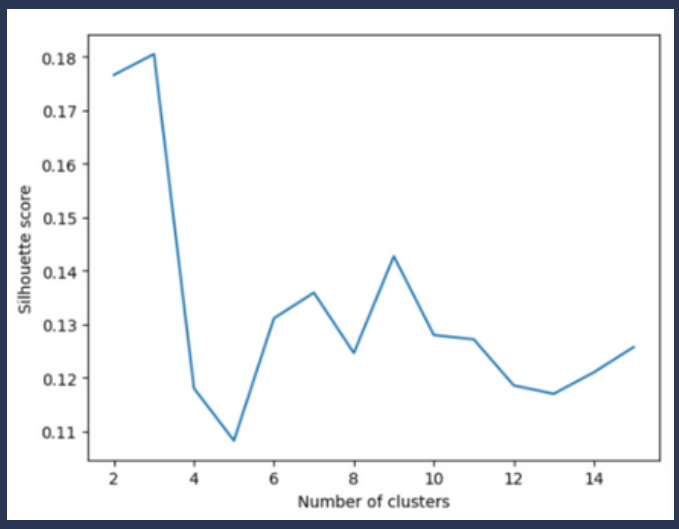


$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

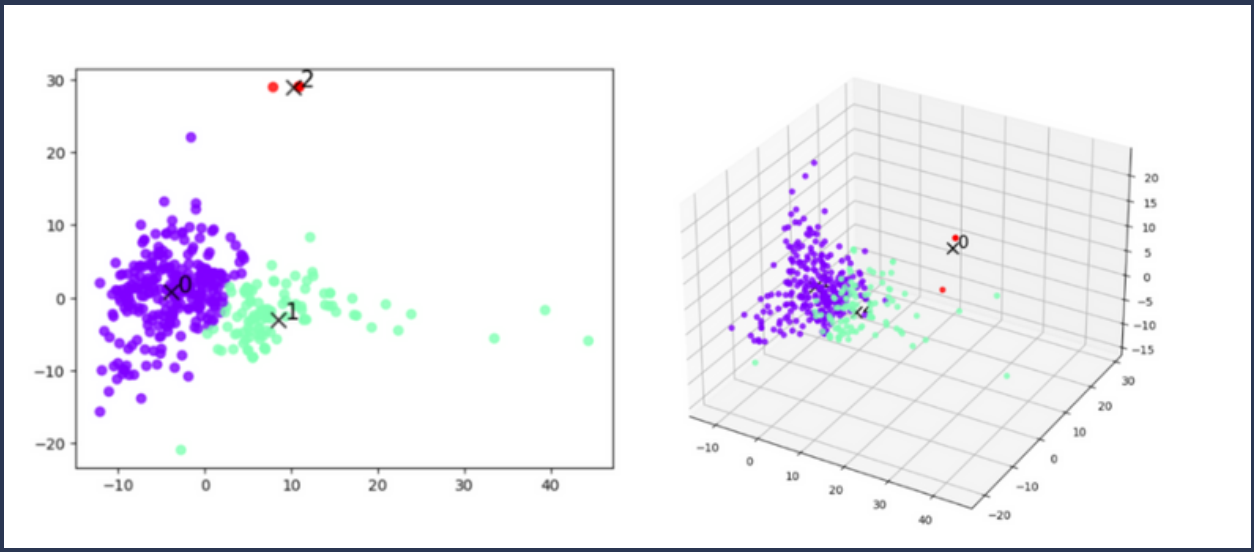
Visualization of inertia as a function of the number of clusters and the formula for inertia used in K-means.

The second method uses the silhouette_score() function from the scikit-learn library. It evaluates the effectiveness of a classification between -1 and 1 (-1 being the worst), measuring the proximity of clusters to each other.

Performance of the K-means algorithm based on the number of clusters.



After this analysis, we will determine an optimal number of clusters based on our objectives. In our example, we choose 3 clusters. However, K-means is a non-deterministic algorithm. Indeed, the algorithm used behind is the Lloyd-Max algorithm, and the first step involves randomly choosing the cluster centers. This simple step does not allow for reproducibility unless a fixed state with a seed is defined. To optimize reproducibility and improve the result, we use a seed to set a fixed initial state. Thus, we obtain the following result:



Visualization of clusters in 2 dimensions and 3 dimensions.

When observing the comments by clusters, we get:

- Cluster 1 (size = 265 comments) : “hydraulique fluviale”, “Hydrologie Hydraulique”, “Génie des procédés Hydraulique urbaine”, “génie des procédés”, ...
- Cluster 2 (size = 118 comments) : “Insertion Pro”, “Excel”, “Communication”, “Ecologie”, ...
- Cluster 3 (size = 5 comments) : “Tous”, “Tous”, “stage”, “Tous”, “Tous”.

This analysis shows that the comments seem to be well grouped according to their ideas. Cluster 3 is quick to analyze due to its small size. However, this is not the case for clusters 1 and 2, each containing over a hundred comments. It would still be too time-consuming to verify this manually.

Generation of cluster descriptions

To speed up the analysis, we use the ChatGPT 4 API to automatically generate a title and description for each cluster. This helps in understanding commonalities within a cluster and quickly differentiating between groups. By examining the number of responses for each cluster, we swiftly gain insight into the students' opinions.

A crucial step is creating the prompt for ChatGPT. Although this part does not pose major difficulty, ensuring quality input is essential for reliable results. We provide a brief description of what we want, a title, a description for each cluster, along with the question graduates responded to. We also specify the expected response format (a JSON) and the necessary attributes, while limiting to a maximum of thirty messages per cluster to avoid information overload.

Once the prompt is created, the request is sent, and the response is associated with each cluster. For the example of useful STE (Science, Technology, Engineering) programs, we obtain the following descriptions:

- Cluster 0
  - Title: Hydraulic, water treatment, and associated software
  - Description: This cluster gathers comments highlighting the importance of courses related to hydraulics, water treatment, and the software used in these fields. Comments also emphasize the value of field trips, internships, and concrete projects to familiarize oneself with the working world and real-world issues.

- Cluster 1
  - Title: Transversal skills and computer tools
  - Description: This cluster compiles comments emphasizing the importance of transversal skills such as communication, project management, labor law, or English. Comments also mention the usefulness of computer tools like Excel, QGIS, PCSWMM, or programming. The comments stress the importance of soft skills and understanding how companies operate.

- Cluster 2
  - Title: All courses or internships
  - Description: This cluster gathers comments that do not focus on a specific course or field but consider that all courses or internships are useful and relevant for professional practice and career integration.

Results

We analyzed three questions using this method: Which courses do you find most useful for your profession and professional integration? Which courses, absent from your education, would have been useful? What advice could you give to students currently in training to choose their end-of-study internship wisely and succeed in their professional integration? We present the most frequent responses to these questions below. Calculations were performed for each field of study, and the percentage indicates the proportion of comments belonging to the cluster.

1. Which courses do you find most useful for your profession and professional integration?

All fields of study: Enseignements lié aux calculs et la conception (30%), Enseignements techniques variés (11%), Enseignements professionnels ou linguistique (14%) ...

IG : Application pratique et savoir-faire (46%), Enseignements informatique et transversale (15%), ...

GBA : Spécialisation et compétences techniques (35%), Diversité et application des enseignements (31%), ...

MAT : Compétences techniques spécifiques (31%), Eco-conception et travail technique (30%), ...

MEA : Système embarquée et électronique (29%), Programmation et projet technique (17%), ...

MI : Enseignement technique et informatique (23%), Enseignement transversal et gestion de projet (17%), ...

STE : Application spécifiques et compétences diverses (67%), Hydraulique et Hydrologie (32%), ...

For the most useful courses, the response varies significantly depending on the field of study. Indeed, the majority of responses are related to technical and specific skills. More general teachings, such as communication, languages, or computer science, are also very common.

2. Which courses, absent from your education, would have been useful?

All fields of study : Outils, technologie et méthodes modernes (49%), Compétence techniques et financières (35%), Aspects managériaux, entrepreneuriaux et personnels (12%), ...

IG : Enseignements liés au code et la sécurité (34%), Enseignements variés et complémentaires (31%), Enseignements plus approfondis ou plus théoriques (20%) ...

GBA : Innovation, environnement et vie professionnelle (69%), management et compétences transversales (21%) ...

MAT : Compétences industrielles et techniques (54%), Expérience pratique et la sensibilisation industrielle (8%), ...

MEA : Informatique, électronique et développement durable (57%), Conception et communication (26%), ...

MI : Sécurité industrielle et durabilité (48%), Compétence spécifique et complémentaire (26%), ...

STE : Besoin en matière spécifique (63%), Compétences techniques (33%), ...

For the absent courses, once again, it is about specific skills related to the respective fields of study. However, there is an overall sentiment of wanting more in-depth courses.

3. What advice could you give to students currently in training to choose their end-of-study internship wisely and succeed in their professional integration?

Toutes les filières : Importance de la passion et la curiosité (31%), Préparation et réflexion à long terme (19%), ...

IG : Stage aligné avec les objectifs professionnels (31%), Stage pour développer ses compétences (18%), ...

GBA : Stage en fonction du secteur d'activité, service et mission (34%), en fonction de ses envies, ses besoins (35%), ...

MAT : S'y prendre à l'avance et choisir son stage en fonction de son projet professionnel (36%), Bien se renseigner sur les entreprises et le métiers (25%), ...

MEA : Être ambitieux, proactif et ambitieux (28%), Faire ce qui plaît et sortir de la zone de confort (24%), ...

MI : Stage en France (38%), Projet professionnel et recherche d'entreprise (28%), ...

STE : Choisir son stage en fonction de ses intérêts et perspectives (45%), Utiliser le réseau des anciens élèves et enseignants (15%), ...

A piece of advice that many graduates give is to choose an internship based on one's desires and future perspective, ensuring that it aligns with one's future project.

Neural Network : PolyChat
Polytech’s very own “ChatGPT”

Wouldn't it be ideal to have a dedicated ChatGPT tailored for Polytech, trained on a customizable textual dataset of our choice, such as the responses from Polytech’s surveys? By engaging with this specialized PolyChatGPT, it could offer insights and responses specifically related to survey results. For instance, you could connect to PolyChat, ask a question like 'Which field of study requires additional Management courses?' and receive a tailored response. This was the concept we initially envisioned.

We conducted thorough research and consulted with our Deep Learning professor. We realized that we lacked sufficient data, computing power, and the required investment to develop our own version of ChatGPT. Nevertheless, we persevered and opted for a less powerful but functional neural network to assist us in analyzing the available data.

We successfully built a neural network and trained four models, which we named:

- **PolyChatA**
- **PolyChatU**
- **PolyChatI**
- **PolyChatR**

The purpose of each model is:

- **PolyChatA:** Answer questions like, 'For which field was a course in management/programming/English/etc. absent and would have been useful?'
- **PolyChatU:** Answer questions like, 'For which field was the course in management/programming/English/etc. useful?'
- **PolyChatI:** Answer questions like, 'For which field was the course in management/programming/English/etc. useless?'
- **PolyChatR:** Answer questions like, 'For which field should the course in management/programming/English/etc. be reinforced or deepened?'

We trained these models on the textual data of the questions, as each response is labeled with the originating field.

We will explain below the features of the models and how we built them.

Tokenization of the text : NLTK and keras

Before designing any neural network, it is essential to proceed with the tokenization of textual data. Tokenizing data involves splitting a text into smaller units called tokens, usually words, to represent them in a structured way. This step is crucial for neural networks as it transforms the text into sequences of integers, making it easier for models to work with numerical data rather than raw text. In our case, we used the tokenizer from NLTK (Natural Language Toolkit), a natural language processing library, to tokenize all textual data. Then, we converted these tokenized data into numerical sequences using the Keras tokenizer, a machine learning library from TensorFlow developed by Google.

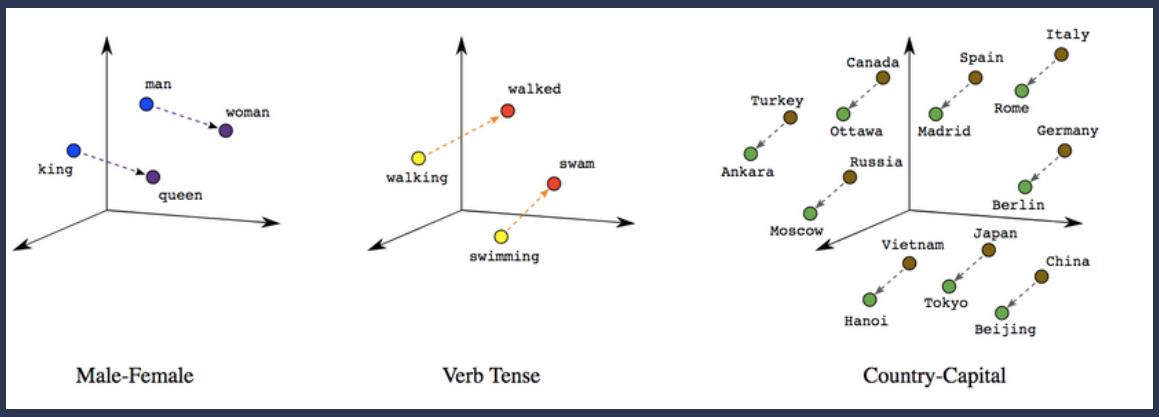
The goal of this step is to prepare textual data for use as inputs to the neural network.

It is important to use the same tokenizer to process new textual inputs once the models are trained. Thus, for each model, we ensured to save the configuration of the tokenizer.

Neural Network with Tensorflow

Our goal was to adhere to the idealistic ultimate goal of Deep Learning, which is to build a network that can adapt to any question. Thus, the final model depends solely on the data and not on the network. It is with this goal in mind that we built our network.

The first layer is an **Embedding** layer, which is used to learn a dense, lower-dimensional representation for the input data. The embedding aims to capture the semantic relationships between words by placing them in a continuous vector space. Each word is represented by an embedding vector, and similar words are positioned close to each other in this space, as we see in this figure:



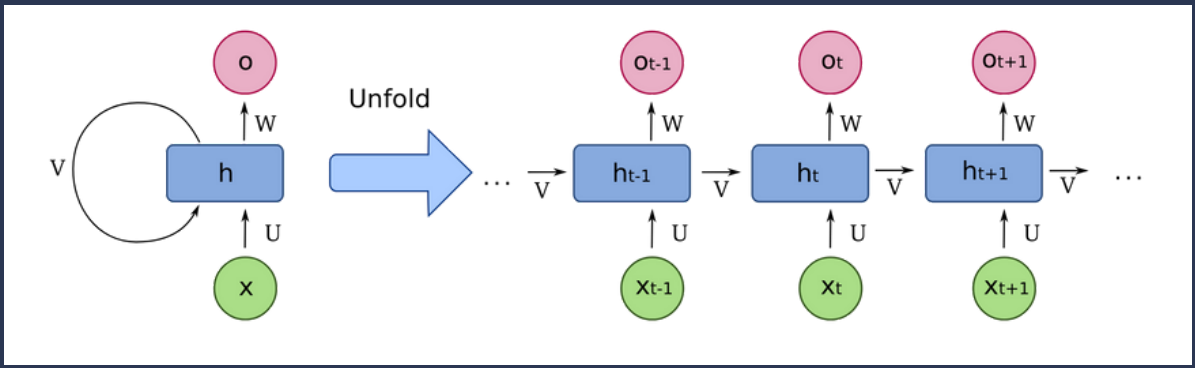
Visualization of the vector space of the embedding

The reasons for which we used this layer are:

- **Dimensionality reduction**
- **Learning relationships:** The embedding captures semantic similarities between words, enhancing the model's ability to generalize.
- **Resource efficiency:** By reducing dimensionality, the embedding layer reduces the number of parameters to learn.

The next layer and one of the most important layers of the network is the **LSTM (Long Short-Term Memory) layer**.

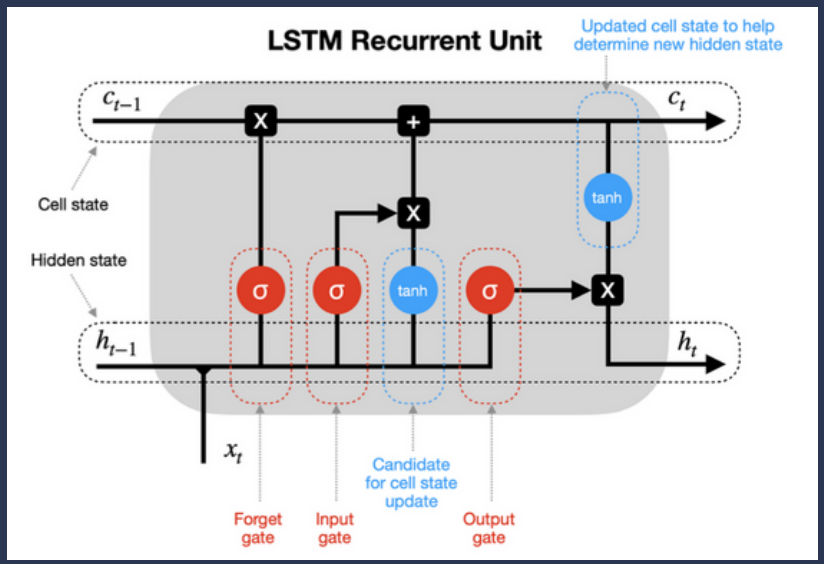
Firstly, an LSTM layer is a **recurrent layer**. Recurrent layers are specifically designed to handle sequential data, maintaining an internal state that takes into account previously processed information in the sequence. This feature makes them ideal for tasks involving sequences, such as natural language processing, which is our case. Here is a clearer diagram:



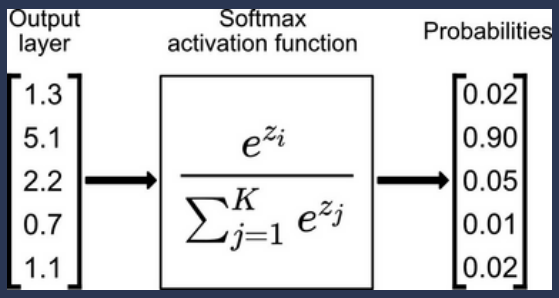
RNN Architecture

However, traditional recurrent layers can pose challenges in learning long-term dependencies, including vanishing gradients, where gradients become too small to have a significant impact, and saturation, where neuron activations reach extreme values.

Hence the invention of **LSTMs**: designed to overcome the long-term dependency issues faced by traditional recurrent layers. LSTMs use gates to regulate the flow of information, allowing them to memorize, forget, or update information based on the context of the sequence, thus solving the vanishing gradient problem. This special architecture enhances the model's ability to learn complex dependencies in data sequences, meaning it can better consider the context from the beginning of a sentence when processing the end of the sentence. Here is a diagram to visualize the gates:



The final layer of the network is a **Dense** layer with a **softmax** activation function. It generates a probability distribution over the different sector classes. Indeed, the softmax function normalizes the scores for each class, transforming the model's outputs into probabilities. Therefore, the class with the highest probability is selected as the model's final prediction for each instance. The softmax function is quite simple:



Hyperparameters

The choice of hyperparameters was done by researching the best practices and also through empirical evaluations.

For the loss function, we chose **Sparse Categorical Crossentropy**, a commonly used metric in training multiclass classification models, particularly when labels are provided as integers, assessing the disparity between the probability distribution predicted by the model and the actual distribution of classes. Sparse Categorical Crossentropy:

$$L_{\text{CE}} = -\sum_{i=1}^{n} t_i \log(p_i), \text{ for n classes,}$$

where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

Sparse Categorical Crossentropy

For the optimizer, we chose **Adam**, which combines adaptive moment estimation techniques and stochastic gradient descent, automatically adjusting the learning rate to accelerate the model's convergence while avoiding the pitfalls associated with manually choosing the learning rate.

We chose **30 epochs, 100 LSTM units,** and a **batch size of 32** after several training sessions to test different values.

Evaluating the models

We split the training dataset into an 80/20 ratio, training the model on 80% of the data and testing it on the remaining 20%. As the main goal was to describe responses, we deliberately chose to optimize the fit on this data.

PolyChatU :

- Number of parameters: **162722**
- Accuracy on the data: **0.9320 - 93%**

PolyChatR :

- Number of parameters: **180226**
- Accuracy on the data: **0.8846 - 88%**

PolyChatA :

- Number of parameters: **180386**
- Accuracy on the data: **0.8783 - 88%**

PolyChatI :

- Number of parameters: **137410**
- Accuracy on the data: **0.7835 - 78%**

Indeed, some models show lower performance than others, and this is due to the fact that those with less satisfactory results have fewer responses, resulting in a smaller amount of data for training.

Results

It is interesting to ask questions to these models, taking into account the results from previous sections. Here are a few questions we asked and their responses:

1. For which field of study should programming be reinforced?

Microelectronique Et Automatique (MEA)

2. Which field of study found maths useful?

Mecanique Structures Industrielles (MSI - apprentissage)

3. For which field was management absent and would have been useful?

Genie Biologique et Agroalimentaires (GBA)

4. Which field of study found the course on law useless?

Materiaux (MAT)

5. Which field found business games useless?

Microelectronique Et Automatique (MEA)

6. Which field found Spanish useless?

Informatique et Gestion (IG)

7. For which field should the maths course be reinforced or deepened?

Genie Biologique et Agroalimentaires (GBA)

8. Which field need the course on python to be reinforced or deepened?

Microelectronique Et Automatique (MEA)

9. Which field found finance to be useless?

Informatique et Gestion (IG)

We can still ask it hundreds of questions, and we invite the decision-makers at Polytech to do so. We find the same results as those we have found in the previous analyses. Obviously, the models are never perfect, and it is always good to cross-check these results with other analyses.

Attempt at BERT construction

We sought to improve the input embedding of our network as well as our tool in general. To achieve this, we undertook the construction of a BERT model, which is a component of the transformer architecture as described in "Attention Is All You Need" written by *Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin* and is widely used today for text embedding. BERT is employed by pre-training a model on large textual corpora, enabling the model to learn rich contextual representations for each word. Then, for a specific task, the BERT model can be fine-tuned on a targeted dataset, adjusting its representations to better fit the particular task, and these representations can be used as contextual embeddings for the text. We were unable to implement BERT due to the complexity of the task.

Conclusion

Throughout this analysis, we have successfully implemented the entire process of studying textual data. We conducted data extractions, transformations, global analyses, and proceeded with more in-depth and refined analysis methods. We presented the results and interpretations throughout this report. Additionally, we provided tools and an interactive application that can be reused and improved when Polytech receives responses from future surveys in the years to come.

Annexes

Code

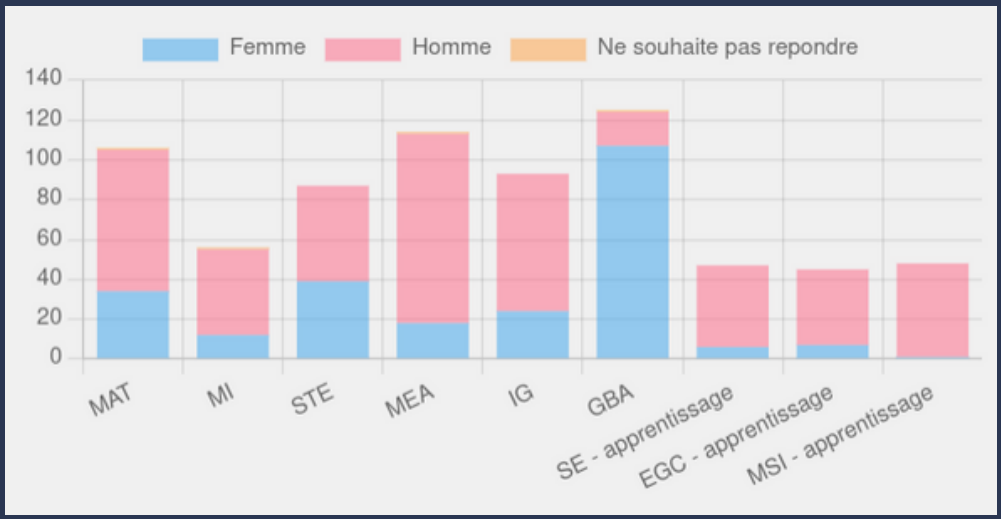
- **ETL, Tools used, backend, etc. :** https://github.com/alexdeloire/data_science_backend
- **Dashboard, chat, frontend :** https://github.com/alexdeloire/data_science_frontend
- **The app is deployed here :** https://poly-analyse.onrender.com/

Examples of Graphs

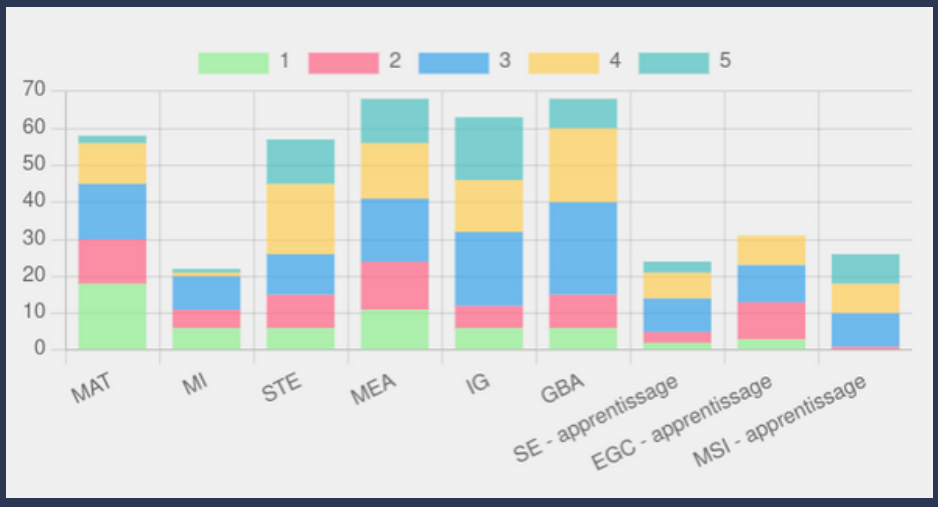
Here are examples of graphs for 2023; for other years, we invite the reader to use our interactive decision tool.

Number of answers by field of study

The breakdown of gender distribution among participants of the survey for each field of study.

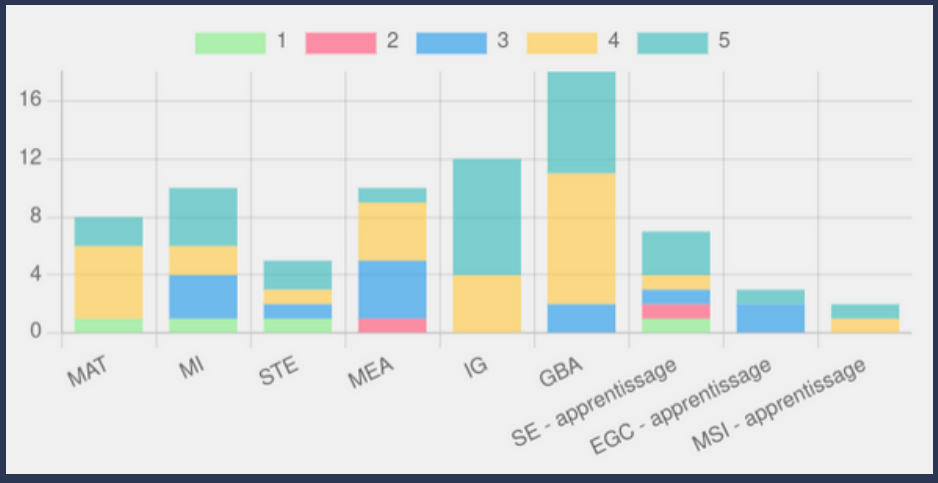


Distribution of responses to the question: Do the following elements seem to have played a role in your recruitment? - the reputation of the training program.

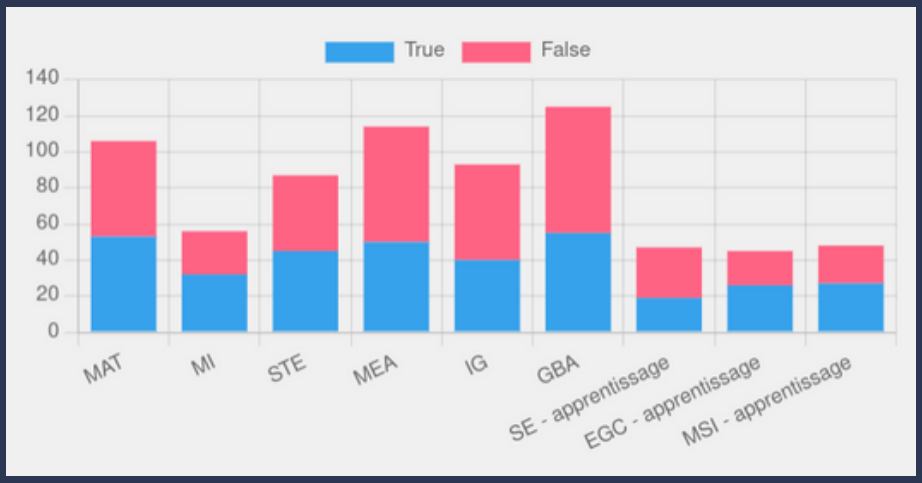


Distribution of responses to the question (True: the person answered, False: the person did not answer): Which courses do you find most useful for your professional practice and your professional integration?

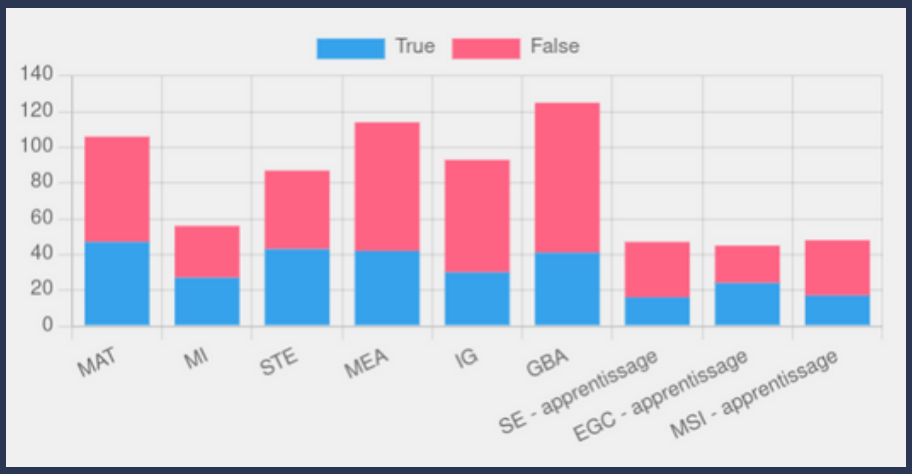
Distribution of responses to the question: Do the following elements seem to have played a role in your recruitment? - the training



Distribution of responses to the question (True: the person answered, False: the person did not answer): Among the courses provided by the school, which ones deserve to be deepened or reinforced?



Distribution of responses to the question (True: the person answered, False: the person did not answer): What courses, absent from your training, would have been useful to you?



Distribution of responses to the question (True: the person answered, False: the person did not answer): What courses, present in your training, do you find useless?