
ethics
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE
ESCUELA DE INGENIERÍA Y CIENCIAS

# EthicApp Project at UChile

## NLP for characterising responses to ethical dilemmas

**Camilo Carvajal Reyes**
**Universidad de Chile**

**3rd November, 2023**

# Outline

Today we will:

- Simulate the experience of facing an ethical dilemma with an interactive activity
- Introduce the concept of dilemmas for evaluating the ethical competency of students at an engineering school
- Present the *EthicApp* tool and its associated dataset
- Past and current natural language processing methods for evaluating ethics and morality in written text
- Future research directions

Repository: https://github.com/camilocarvajalreyes/ethicapp-nlp

# About myself

**Camilo Carvajal Reyes**
dim.uchile.cl/~ccarvajal/ - ccarvajal@dim.uchile.cl

- MSc candidate in Data Science @ U. de Chile
  Current research: safeness for score-based generative models
  Supervised by Felipe Tobar and Joaquín Fontbona
  MDS7203 Deep Generative Models: Teaching assistant
- Mathematical Engineering @ U. de Chile
- Engineering degree @ CentraleSupélec, France
  Research track: word embeddings and semantics of LLMs
- Co-founder of AEDIA: Association for Ethics in Data and Artificial Intelligence.
- Language models for software logs: transformer-based solutions for error analysis at the European Southern Observatory.

# Our team

Ethics Unit (*Unidad de Ética*) at the Faculty of Mathematical and Physical Sciences U. de Chile. Part of the Department of Transversal Studies in Humanities for Engineering (ETHICS).

- Josefa Cerda Maureira
- Pablo Ramirez Rivas
- Assistants: Eduardo Hurtado Mila & me.

# Table of Contents

# Table of Contents

# The Adela case

In Chile, vitamin D deficiency is a serious problem in both older adults and children.

A group of professionals found an ancestral fruit of the Diaguita communities with a high concentration of vitamin D and an attractive taste for consumption.

**Adela**, an engineer of the team, designs the production process of a new food based on this fruit. However, she faces challenges, as the tree only grows near rivers and needs abundant sunlight, which makes it difficult to bring it to more southern areas, so it must be grown in the drought-stricken Norte Chico region.

# The Adela case

In addition, to conserve vitamin D during transport to other areas, the team decided to freeze-dry the fruit and add preservatives. Although they are not yet legally required to integrate the Diaguita communities into the project, Adela listens to their concerns about how these changes would affect their traditions.

Despite the changes being necessary to maintain the benefits of the product, the procedure goes against the communities' traditional practices, which are a fundamental part of their identity.

# The Adela case

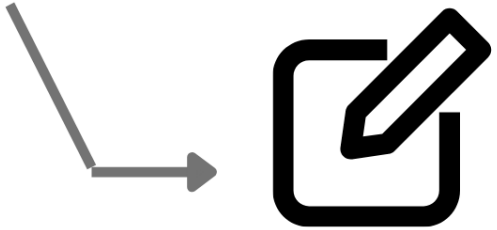
**Considering the population that will benefit from this new food and how the production process will affect the identity traditions of the indigenous community, Adela should :**

*Produce the food against the vitamin deficit*

1 2 3 4 5 6

rating scale

*Protect the identity traditions*



**Text justification**

**Your turn to participate in the activity**:

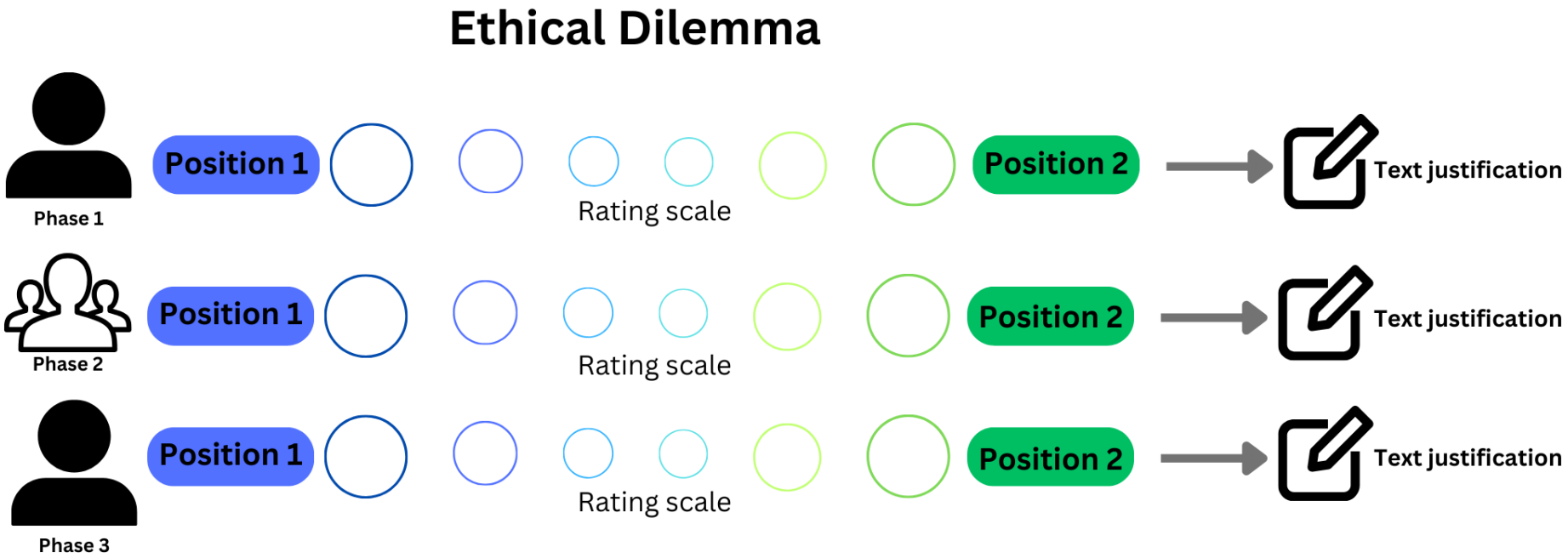
Link to the poll

# The Adela case

**Your turn to participate in the activity!**

Link to the poll

# The Adela case




**Now time for a group discussion**:

Link to the poll

# Table of contents

# Table of Contents

# Ethical formation

"Ethics consists of the ability to provide a reasoned and reasonable basis for one's own actions and for professional actions in particular, within a framework of principles and values that are intended to be universal because they are widely accepted." [1]

# The context of U. de Chile

The Ethics Unit is in charge of developing the ability of ethical commitment at the U. de Chile engineering school.

- Approximately 1000 students in each cohort
- The evaluation and teaching take place on the first three semesters, during the courses from the Innovation Area.
- Desired competence: “Reflect on one’s own actions and their consequences, within the framework of honesty, responsibility and respect, seeking excellence and rigour in their actions in academic contexts, in interpersonal relationships and with their environment”.
- Evaluation indicators include the principles of:
  - responsibility,
  - respect,
  - integrity.

# Ethical Dilemmas

The Ethics Unit makes use of ethical dilemmas for the evaluation of competencies [2]:

Moral dilemmas are stories in the face of which a person must **choose between two possible courses of action**, both of which carry with them a positive charge and a negative one.

Generally, dilemmas rest, as far as their valuation is concerned, on the consequences and/or impacts of the decision to be taken. The story that contains the dilemma can be real or plausible, serving both for ethical formation [3].

# Ethical Dilemmas

There are several models of moral dilemmas [4], but what they all have in common are the following elements:

1. They have two defined positions
2. Each of these contains positive and negative aspects and/or consequences
3. The need to make a decision is imposed.

# Table of Contents

# Use of NLP in morality assessment

## Language models reflecting on human morality

There are several works that seek the prediction/identification of moral elements in natural text. In this section we will address some of them. First, Garten et al. 2016 seeks the automatic detection of moral rhetorics [5]. For this they use the dictionary of moral foundations [6] (from the Theory of Moral Foundations TMF [7]), combined with word embeddings. Another similar work is that of Xie et al. who evaluate models in the classification of moral dilemmas in the different foundations, thanks to which they conclude that language models have an advantage over models such as distributional representations [8].

TMF: (Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, Sanctity/Degradation, Liberty/Oppression)

# Use of NLP in morality assessment

## Morality studies using NLP

On the other hand, Kennedy et al. seek to predict an individual's own moral concerns using evidence of moral language written by the individual (Facebook statuses of users who have answered the moral foundations questionnaire). Different language processing techniques were used to predict the scores obtained by users, for each of the moral dimensions posed in the TMF [7]. The variety of text vectorization methods tested, including *latent dirichlet allocation* (LDA) [10], *word embeddings*, DFM occurrence counting [6] and BERT [11], which corresponds to a deep language model. It is the latter that obtains better results. Finally, both dictionary counts and LDA were used to interpret which specific linguistic elements explained each foundation separately.

# Table of Contents

# The application

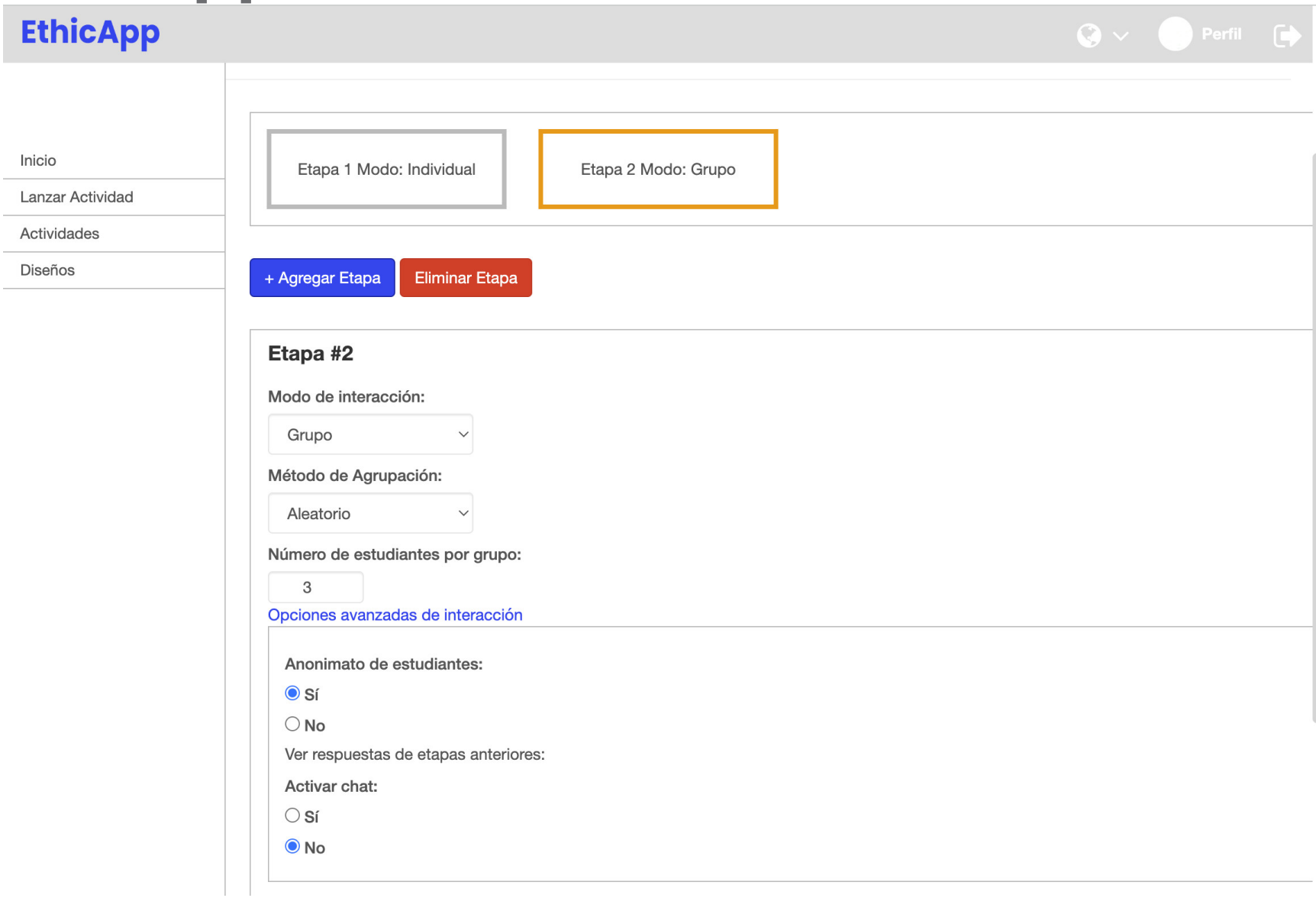
*EthicApp* consists of an application[1] developed at the Economics and Business Faculty of U. Chile and the Engineering Faculty of U. Andes. The tool allows for the implementation and monitoring of ethical dilemmas [12].

Claudio Álvarez
Director de Ingeniería y BDFL
Prof. Asociado, Fac. Ingeniería
Universidad de los Andes, Chile
calvarez (at) uandes.cl

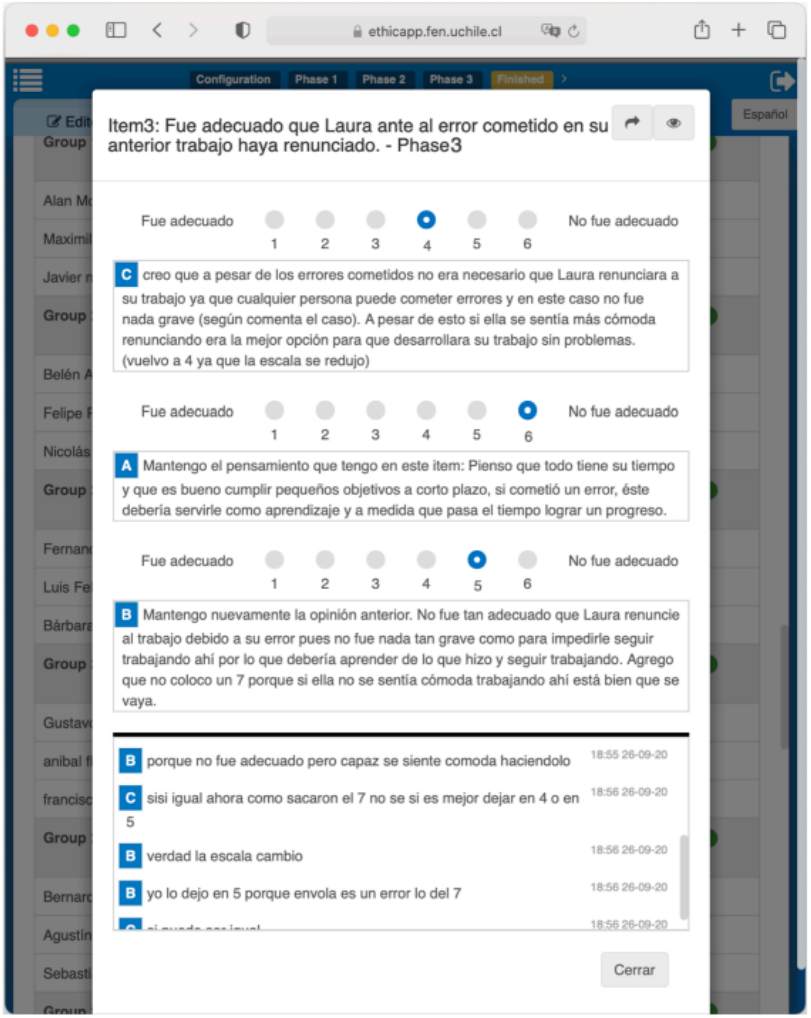
Gustavo Zurita
Director Científico
Profesor Titular, DCS
Universidad de Chile
gzurita (at) fen.uchile.cl

[1] More information can be found on: https://www.ethicapp.info/.

# The application

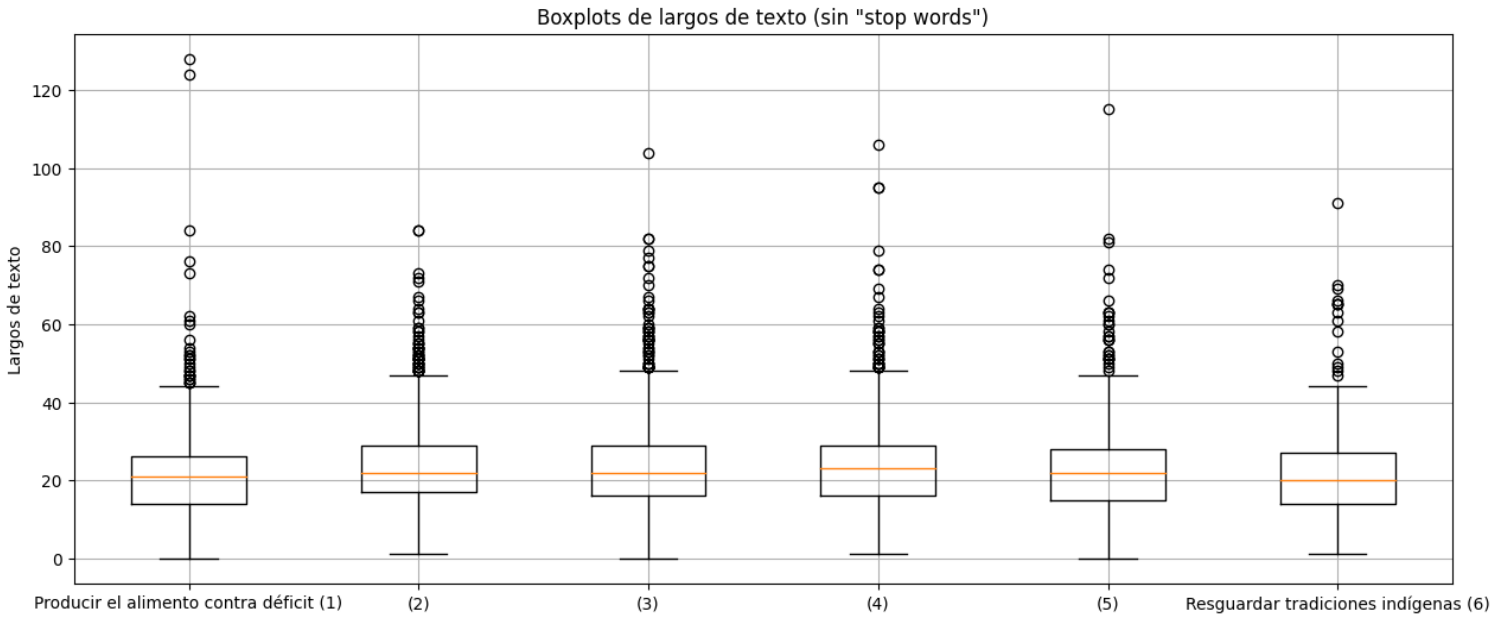


# The application

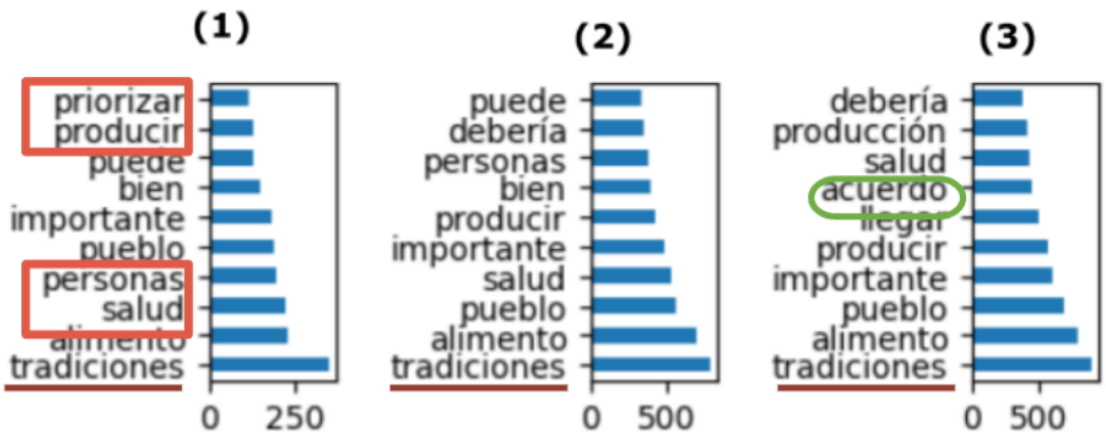
# Data characteristics

| Case | Courses | No. students | No. groups |
|---|---|---|---|
| Julieta | 1 | 819 | 247 |
| **Adela** | **3** | **1866** | **515** |
| Laura | 1 | 602 | 335 |
| Alicia | 2 | 1628 | 549 |

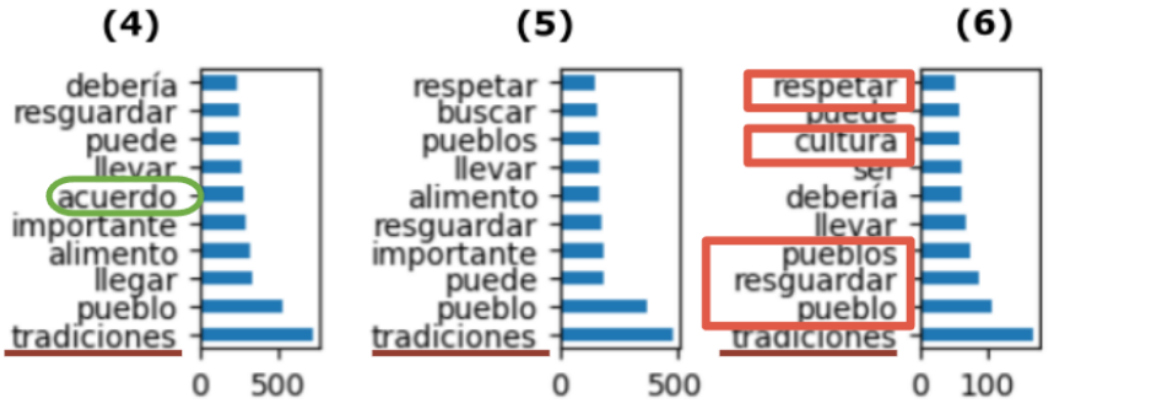


# Common tokens

**Produce the food against the vitamin deficit**




**Protect the identity traditions**



# Frequency of selections

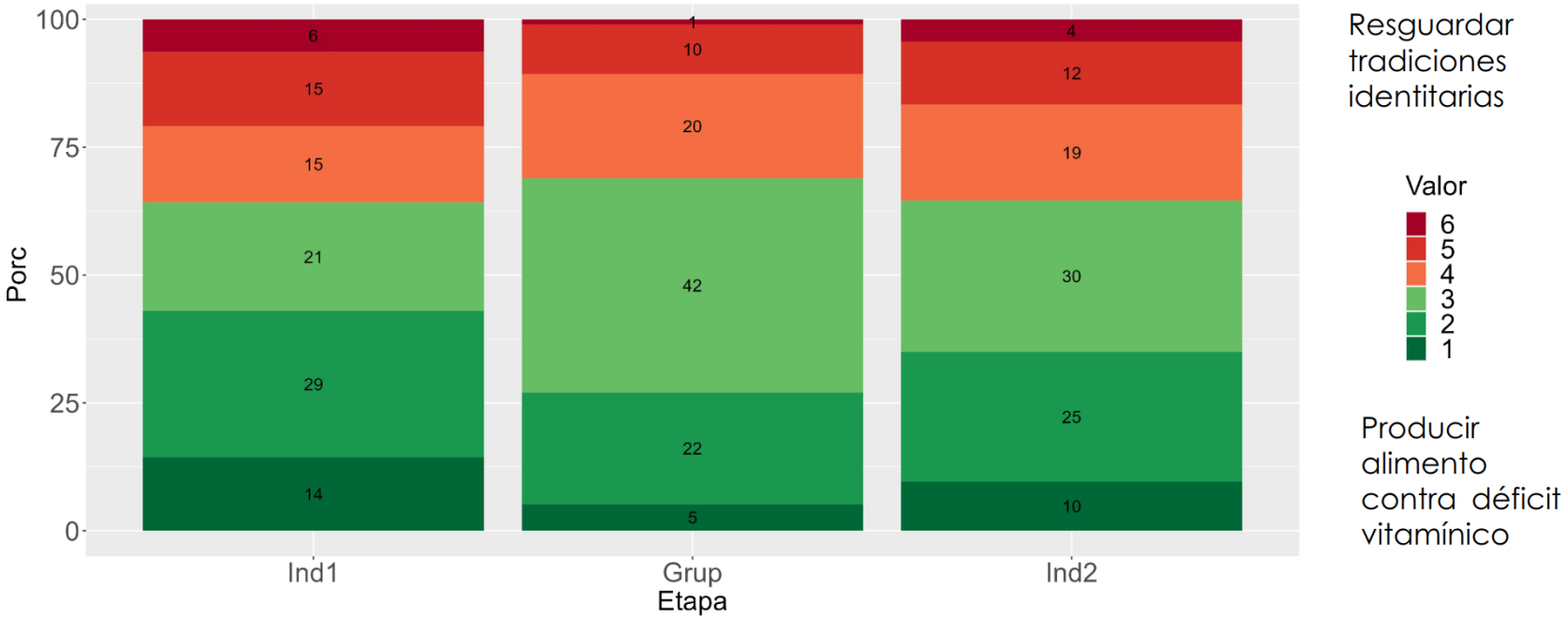
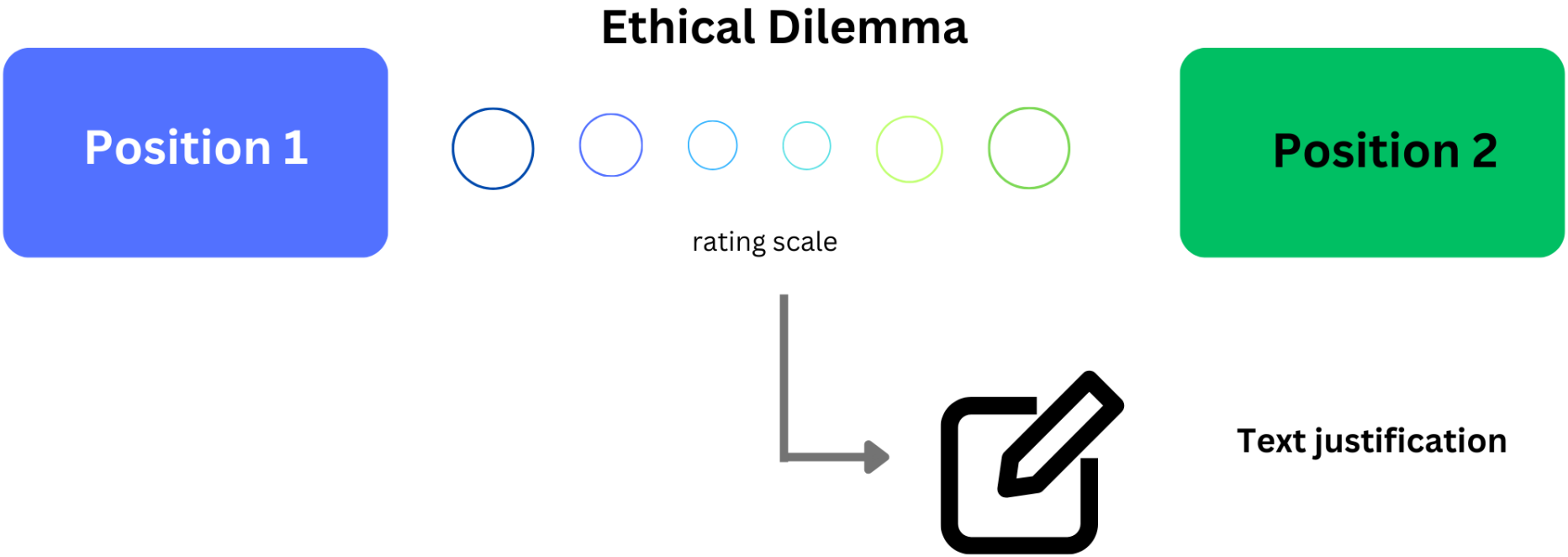



Figure: Students' choice on the Adela dilemma for a course section on each phase.

# Table of Contents

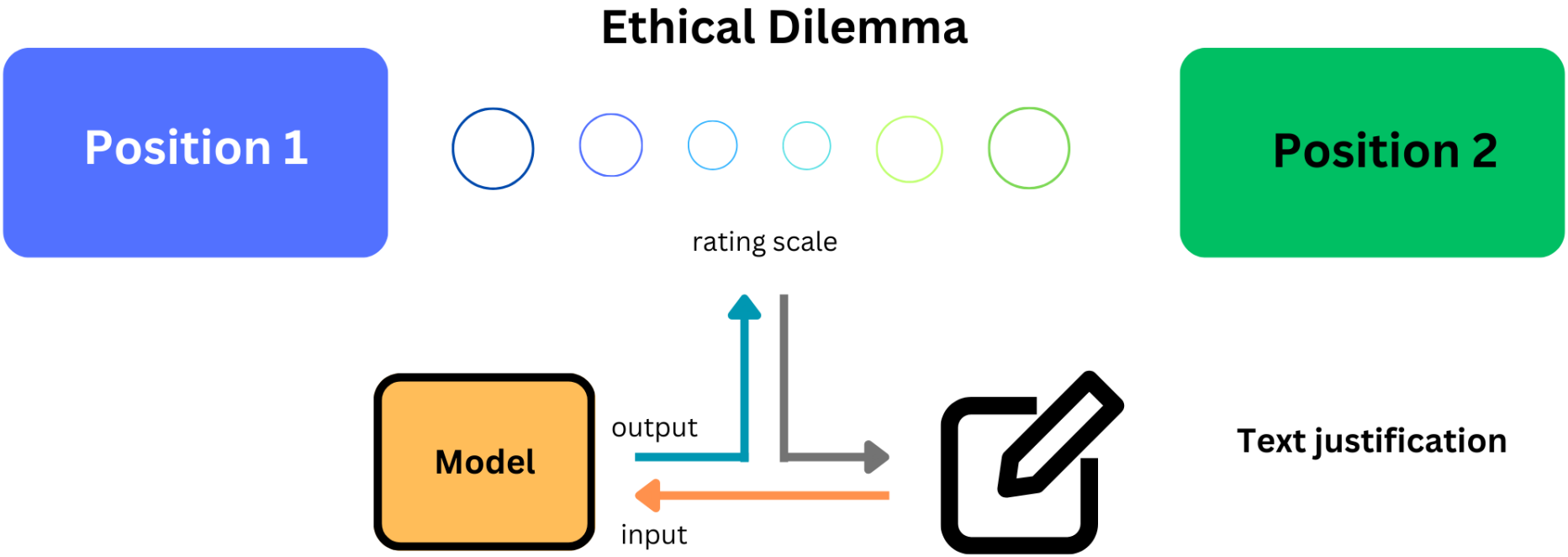
# Predicting students answers

We propose the use of models for predicting the assessment of the problem, using the text. Doing this with interpretable models will give us an idea of what **semantic elements** were used to choose such an option. This complements the work from the *EthicApp* development team, which are using deep learning techniques in a follow-up work from their first NLP-based assistance tool [13].



# Predicting students answers

We propose the use of models for predicting the assessment of the problem, using the text. Doing this with interpretable models will give us an idea of what **semantic elements** were used to choose such an option. This complements the work from the *EthicApp* development team, which are using deep learning techniques in a follow-up work from their first NLP-based assistance tool [13].



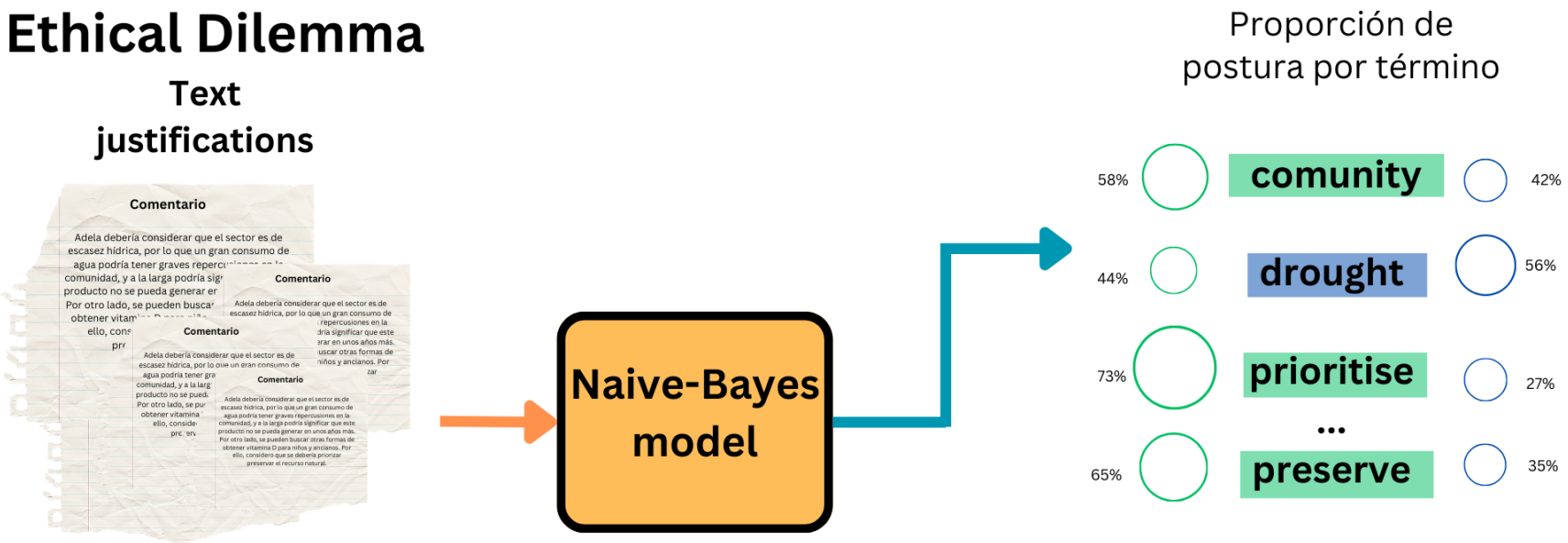
# Our approach: interpretable models

During the first stage of the project, we give priority to count-based methods: Naive-Bayes and LDA since:

- They are interpretable,
- They allow us to perform a token-level analysis,
- They are extremely lightweight when compared to deep-learning models.

# Our approach: interpretable models

The supervised Naive-Bayes model, which naturally assigns probabilities per position to each word, is a classification model that will be tested by separating the six positions as independent cases, as well as grouping the options in two. By analysing the probabilities, it is possible to identify those elements that contribute more to the final result of the algorithm, obtaining the interpretative capacity.



# Preliminary findings

The following tokens characterise each choice, allowing us to "take a broad picture" of the common arguments.

| Produce the food against the vitamin deficit | | | | Protect the identity traditions | | | |
|---|---|---|---|---|---|---|---|
| **Word** | **Prob** | **Word** | **Prob** | **Word** | **Prob** | **Word** | **Prob** |
| **1. *salvar*** | 0.985 | **6.** *poniendo* | 0.942 | **1.** *siglos* | 0.880 | **6.** *usuario* | 0.761 |
| **2. *juego*** | 0.958 | **7. *riesgo*** | 0.938 | **2. *suplementos*** | 0.863 | **7.** *consentimiento* | 0.761 |
| **3. *vidas*** | 0.952 | **8.** *tiempos* | 0.934 | **3. *sol*** | 0.825 | **8.** *integridad* | 0.730 |
| **4. *tribu*** | 0.950 | **9.** *cambiarían* | 0.934 | **4.** *existen* | 0.822 | **9. *sacar*** | 0.724 |
| **5. *religiosas*** | 0.948 | **10.** *ayudarían* | 0.934 | **5.** *única* | 0.770 | **10. *obtener*** | 0.723 |

An initial analysis of representative tokens and students' preferences shows that the principle of responsibility prevails over the principle of respect.

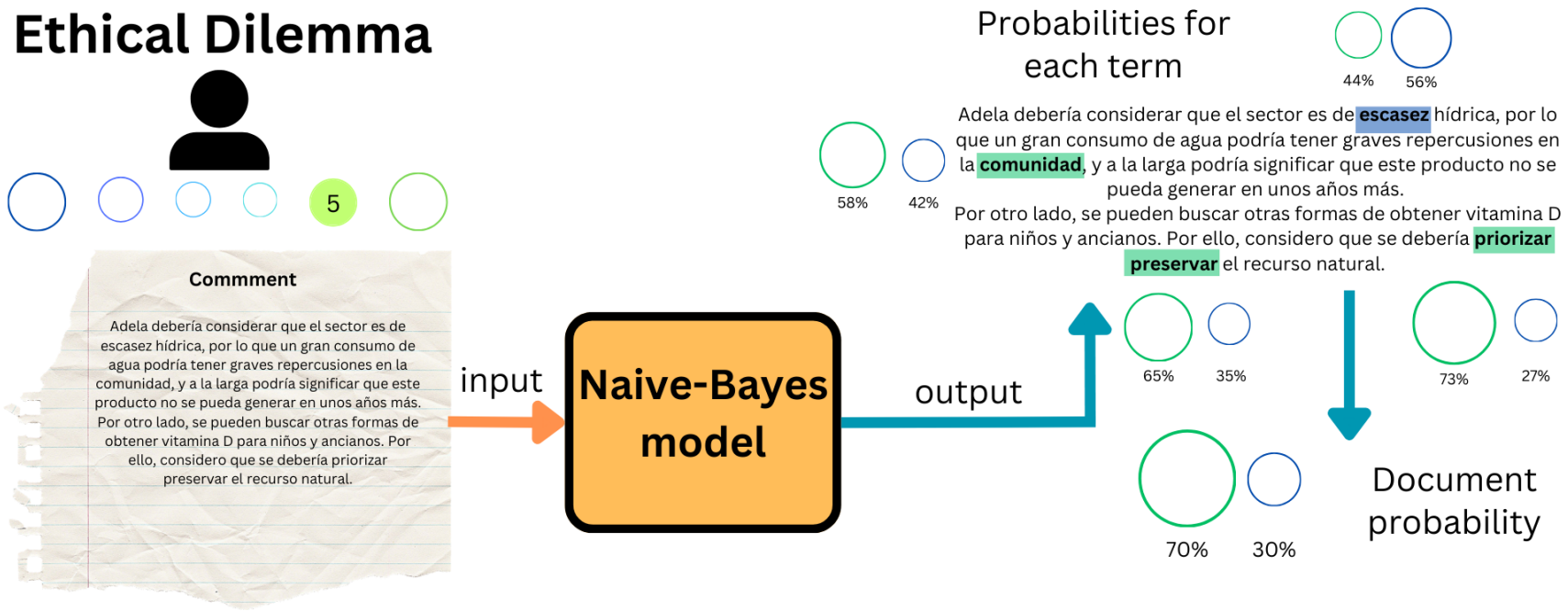
# Visualisation of token probabilities/logits

We make use of the probabilities assigned by the models in order to assist the assessment of written justification. The procedure, inspired by named entity recognition, simply maps the probabilities of our model to a pre-determined colour map.

It is generalisable to other methods by replacing the probabilities with:

- the model weights in linear models
- the feature importance for ensemble models
- topic probabilities when considering topic modelling

# Visualisation of token probabilities/logits



# Visualisation of token probabilities/logits

In this example (beta version):

- We take a simple Naive-Bayes model trained with monograms.
- The bluer-darker the token, the more probability it has of being used in an argument to protect the identity traditions.

```
visualise("""La forma en que está diseñada la innovación nos parece bastante nociva, y que de todas maneras es
          modificable, de modo de producir el producto de manera respetuosa con las comunidades y los ecosistemas.
          Es una buena herramienta y por tanto se debería proseguir pero adaptando la forma""")
```

[10] ✓ 0.0s Python

...

La forma en que está diseñada la innovación nos parece bastante nociva, y que de todas maneras es modificable, de modo de producir el producto de manera respetuosa con las comunidades y los ecosistemas. Es una buena herramienta y por tanto se debería proseguir pero adaptando la forma

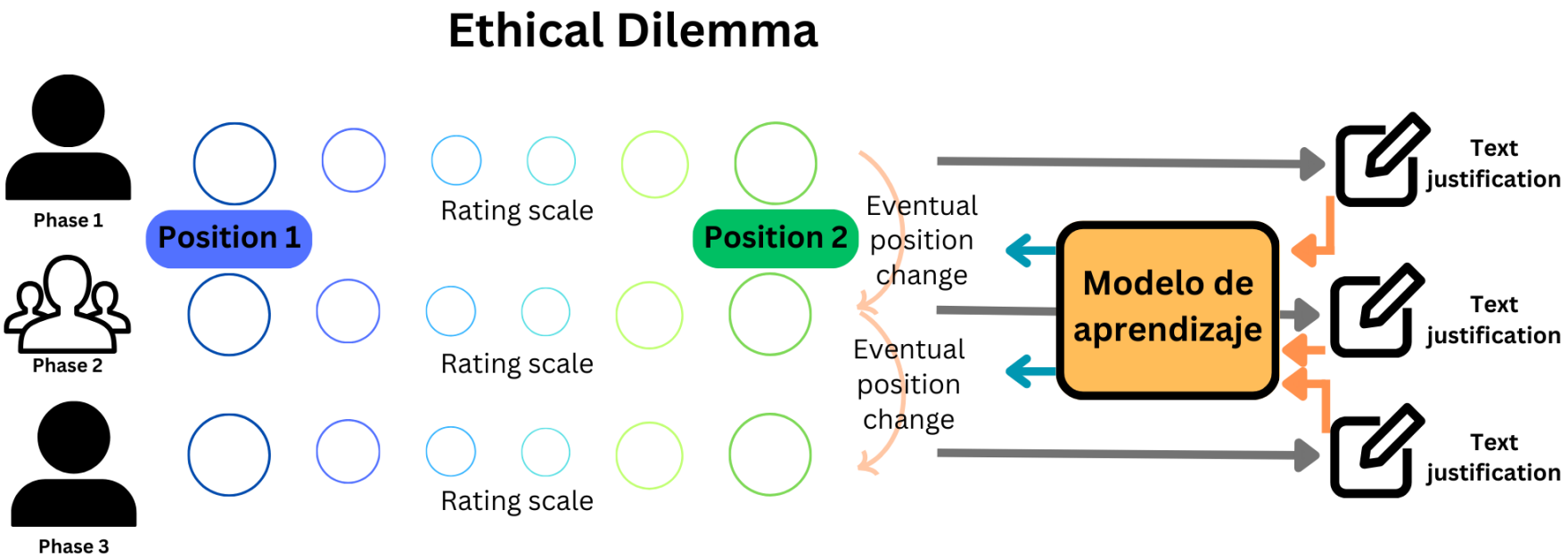
# Topic Modelling

The use of topic modelling is particularly challenging since one could argue that the main topic is common for all text justifications. This time *BERTopic* performs better (qualitatively) than the LDA model. Most notable, it is able to identify some outlier topics.

| Topic | Key words | Prob | Example |
|---|---|---|---|
| 0 | **food, traditions, people** | 70,99% | Adela should produce the food. I think it is more important to benefit people who have a vitamin deficit, because it is important to preserve the traditions of identity. But that is not about the health of the population. |
| 1 | **traditions, people, agreement** | 16,8% | First it would be to try to reach an agreement with the people, but if it is not possible, if it can save people this fruit then it is more important the life of the people than a tradition. |
| 2 | **health, important, persons** | 5,27% | I consider it much more important to prioritize the majority of the population for the sake of health. |
| 3 | **project, innovation, if** | 2,14% | Although Adela has no legal limitations for her innovation project, she should take into account the opinion of the native peoples of the area to try to reach an agreement with them, in order to respect their traditions and culture. |
| 4 | **Chile, filling, chilean** | 1,83% | Hopefully, as far as possible, an agreement can be reached with the native peoples to respect even a part of it. But definitely producing food generates a greater good since it can help more than 50% of the children and older adults in Chile. |
| 5 | **opinion, maintain, no** | 0,97% | I maintain my position |

# Predicting changes in position

Currently, a team from the Master of Data Science programme works on the specific problem of modelling the changes in position.




They also have access to the group chats, which were used to anonymously decide on a common position.

# Table of Contents

# Incorporating interpretable DL

We do not rule out the use of techniques that make deep learning models interpretable.

One such example is *cockatiel* [14], which is designed to generate explanations with concepts for neural network classifiers.

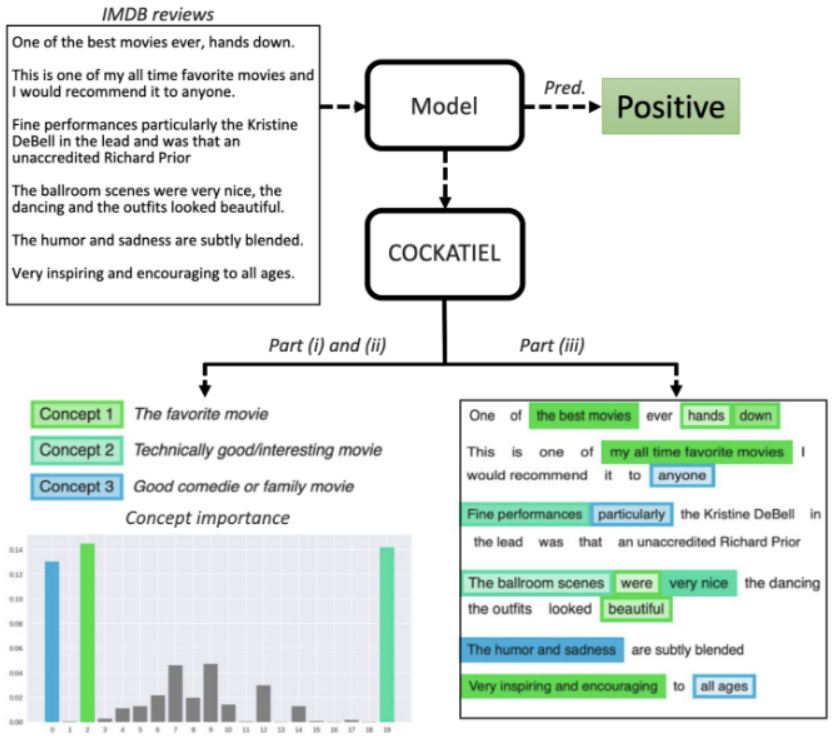
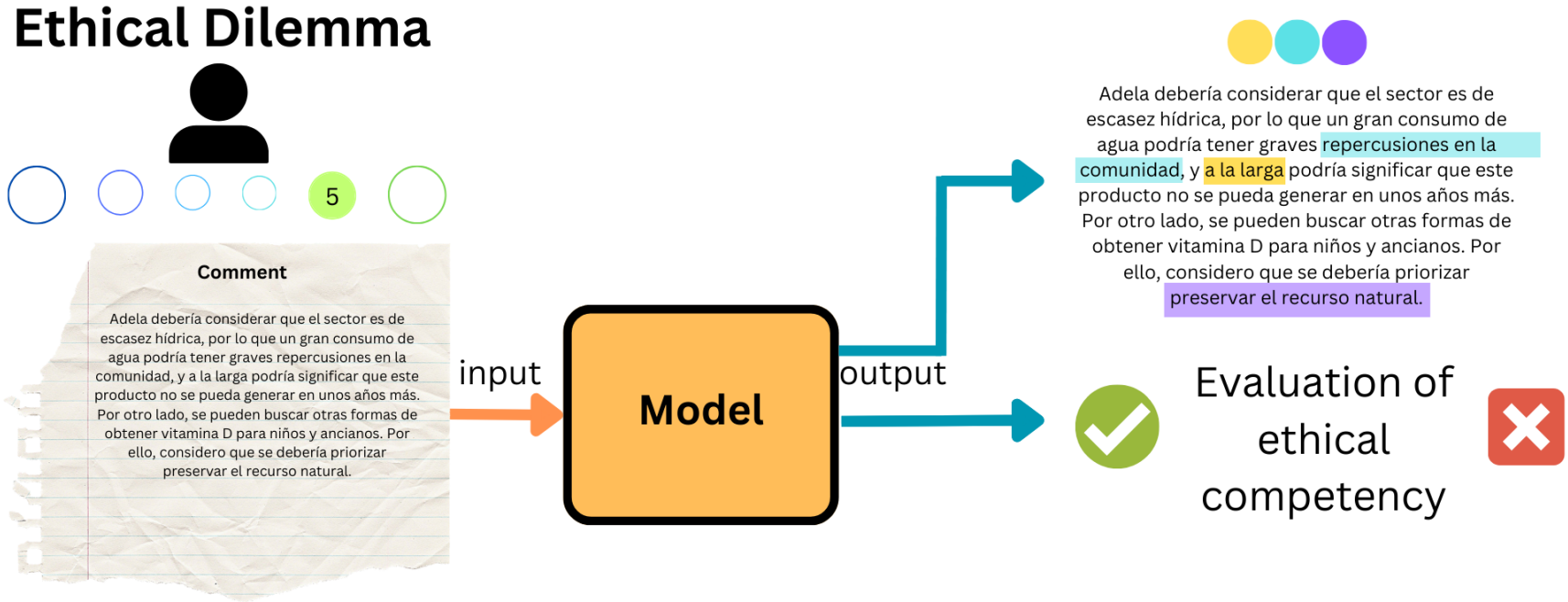



Figure 1: **An illustration of COCKATIEL**. Given some sentences of IMDB reviews, COCKATIEL *(i)* identifies concepts for prediction, *(ii)* ranks them, and *(iii)* gives the most important elements for each concept (to help us interpret the concept).

# Dataset expantion

It is considered for the future the development of a deep learning method that is capable of stating those variables (desirably, concepts) that are important in reflecting ethical competence. For example, detecting specific indicators such as responsibility, respect and integrity.

Furthermore, an expansion of the dataset with expert evaluation is a prospect we are also considering.



# References I


[1] Ramírez Rivas, P. (2012). Formación ética en Ingeniería. Reflexiones y desafíos. Fraternidad y educación: un principio para la formación ciudadana y la convivencia democrática, 63-91.

[2] Ramírez Rivas, P., Guerrero, S., Cerda Maureira, J., Ross, J. P., Flores Mandeville, G. (2022). La formación ética canalizada mediante la tecnología. Experiencia y resultados preliminares del uso de la herramienta web Ethicapp. XXXIV Congreso Chileno de Educación en Ingeniería.

[3] Meza Rueda, J. L. (2008). Los dilemas morales: una estrategia didáctica para la formación del sujeto moral en el ámbito universitario. Actualidades pedagógicas, 1(52), 13-24.

[4] Ruíz-Cano, J., Cantú-Quintanilla, G. R., Ávila-Montiel, D., Gamboa-Marrufo, J. D., Juárez-Villegas, L. E., de Hoyos-Bermea, A., ... Garduño-Espinosa, J. (2015). Revisión de modelos para el análisis de dilemas éticos. Boletín médico del hospital infantil de México, 72(2), 89-98.

[5] Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., Dehghani, M. (2016). Morality Between the Lines: Detecting Moral Sentiment In Text. Proceedings of IJCAI 2016 Workshop on Computational Modeling of Attitudes.

[6] Graham, J., Haidt, J., Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of Personality and Social Psychology, 96, 1029–1046. https://doi.org/10.1037/a0015141

[7] Haidt, J. (2007). The New Synthesis in Moral Psychology. Science, 316(5827), 998–1002. https://doi.org/10.1126/science.1137651

[8] Xie, J. Y., Hirst, G., Xu, Y. (2020). Contextualized moral inference (arXiv:2008.10762). arXiv. https://doi.org/10.48550/arXiv.2008.10762

[9] Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J., Dehghani, M. (2021). Moral concerns are differentially observable in language. Cognition, 212, 104696. https://doi.org/10.1016/j.cognition.2021.104696

[10] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3(null), 993–1022.

# References II

[11] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Alvarez, C., Zurita, G., Hasbún, B., Peñafiel, S., Pezoa, Á., Alvarez, C., Zurita, G., Hasbún, B., Peñafiel, S., Pezoa, Á. (2021). A Social Platform for Fostering Ethical Education through Role-Playing. In Factoring Ethics in Technology, Policy Making, Regulation and AI. IntechOpen. https://doi.org/10.5772/intechopen.96602

[13] Alvarez, C., Zurita, G., Carvallo, A., Ramírez, P., Bravo, E., Baloian, N. (2021). Automatic content analysis of student moral discourse in a collaborative learning activity. In Collaboration Technologies and Social Computing: 27th International Conference, CollabTech 2021, Virtual Event, August 31–September 3, 2021, Proceedings 27 (pp. 3-19). Springer International Publishing.

[14] Jourdan, F., Picard, A., Fel, T., Risser, L., Loubes, J. M., Asher, N. (2023). COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks (arXiv:2305.06754). arXiv. https://doi.org/10.48550/arXiv.2305.06754


ethics
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE
ESCUELA DE INGENIERÍA Y CIENCIAS

# EthicApp Project at UChile

## NLP for characterising responses to ethical dilemmas

**Camilo Carvajal Reyes**
**Universidad de Chile**

**3rd November, 2023**