



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

- 1 Outline**
- 2 Facing the ethical dilemma: The Adela case**
- 3 Evaluating the ethical competency**
 - Ethical Formation
 - The use of dilemmas
- 4 Use of NLP in morality assessment**
- 5 EthicApp: tool and dataset**
 - The application
 - Data characteristics
- 6 Text analysis using dilemmas**
 - Predicting students answers
 - Predicting changes in position
 - Future work
- 7 References**

Table of Contents

1 Outline

2 Facing the ethical dilemma: The Adela case

3 Evaluating the ethical competency

- Ethical Formation
- The use of dilemmas

4 Use of NLP in morality assessment

5 EthicApp: tool and dataset

- The application
- Data characteristics

6 Text analysis using dilemmas

- Predicting students answers
- Predicting changes in position
- Future work

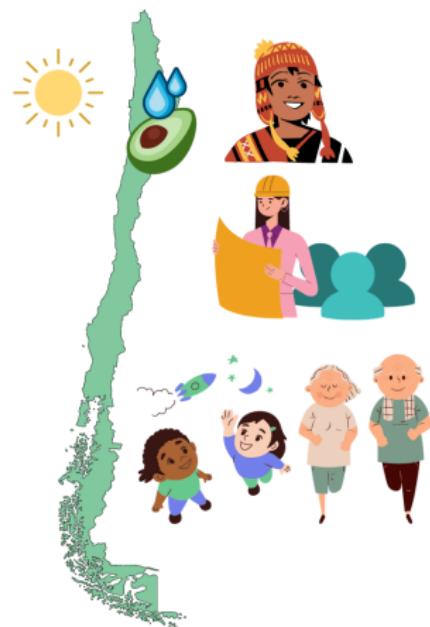
7 References

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

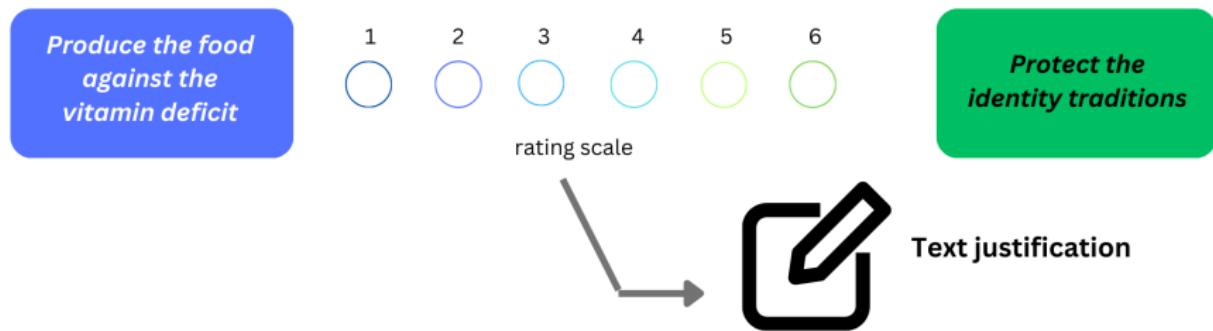
In addition, to conserve vitamin D during transport to other areas, the team decided to freeze-dry the fruit and add preservatives. Although they are not yet legally required to integrate the Diaguita communities into the project, Adela listens to their concerns about how these changes would affect their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



Your turn to participate in the activity:
[Link to the poll](#)



The Adela case

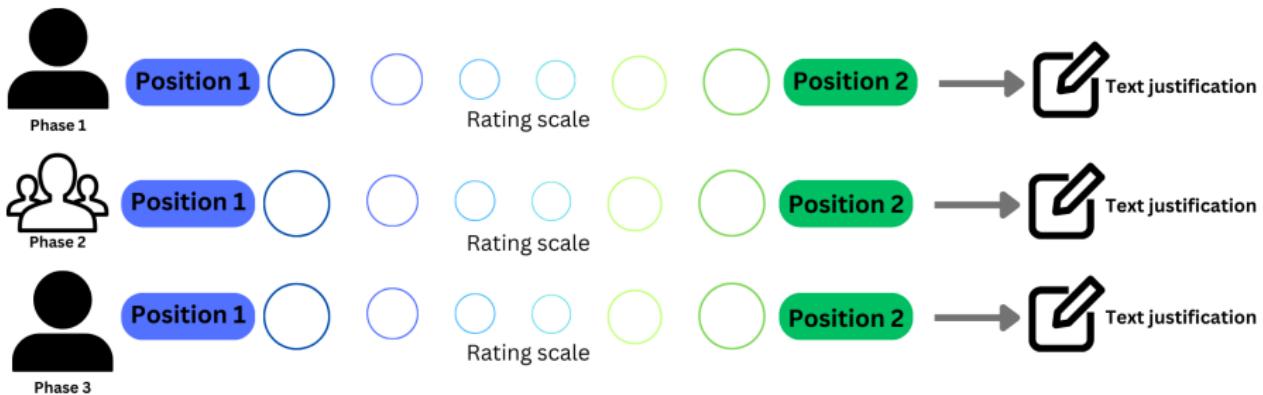
Your turn to participate in the activity!

[Link to the poll](#)



The Adela case

Ethical Dilemma



Now time for a group discussion:
[Link to the poll](#)



Table of contents

- 1 Outline**
- 2 Facing the ethical dilemma: The Adela case**
- 3 Evaluating the ethical competency**
 - Ethical Formation
 - The use of dilemmas
- 4 Use of NLP in morality assessment**
- 5 EthicApp: tool and dataset**
 - The application
 - Data characteristics
- 6 Text analysis using dilemmas**
 - Predicting students answers
 - Predicting changes in position
 - Future work
- 7 References**

Table of Contents

1 Outline

2 Facing the ethical dilemma: The Adela case

3 Evaluating the ethical competency

- Ethical Formation
- The use of dilemmas

4 Use of NLP in morality assessment

5 EthicApp: tool and dataset

- The application
- Data characteristics

6 Text analysis using dilemmas

- Predicting students answers
- Predicting changes in position
- Future work

7 References

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.

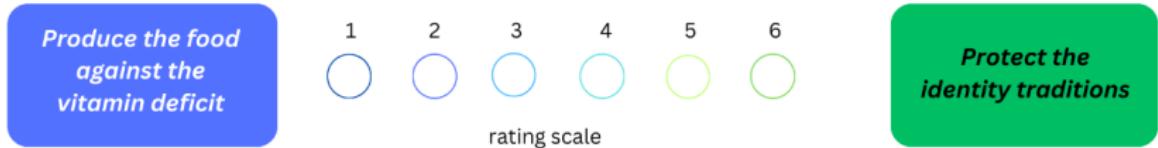


Table of Contents

- 1 Outline**
- 2 Facing the ethical dilemma: The Adela case**
- 3 Evaluating the ethical competency**
 - Ethical Formation
 - The use of dilemmas
- 4 Use of NLP in morality assessment**
- 5 EthicApp: tool and dataset**
 - The application
 - Data characteristics
- 6 Text analysis using dilemmas**
 - Predicting students answers
 - Predicting changes in position
 - Future work
- 7 References**

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

- 1 Outline**
- 2 Facing the ethical dilemma: The Adela case**
- 3 Evaluating the ethical competency**
 - Ethical Formation
 - The use of dilemmas
- 4 Use of NLP in morality assessment**
- 5 EthicApp: tool and dataset**
 - The application
 - Data characteristics
- 6 Text analysis using dilemmas**
 - Predicting students answers
 - Predicting changes in position
 - Future work
- 7 References**

The application

EthicApp consists of an application¹ 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 BD&L
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

¹More information can be found on: <https://www.ethicapp.info/>.

The application

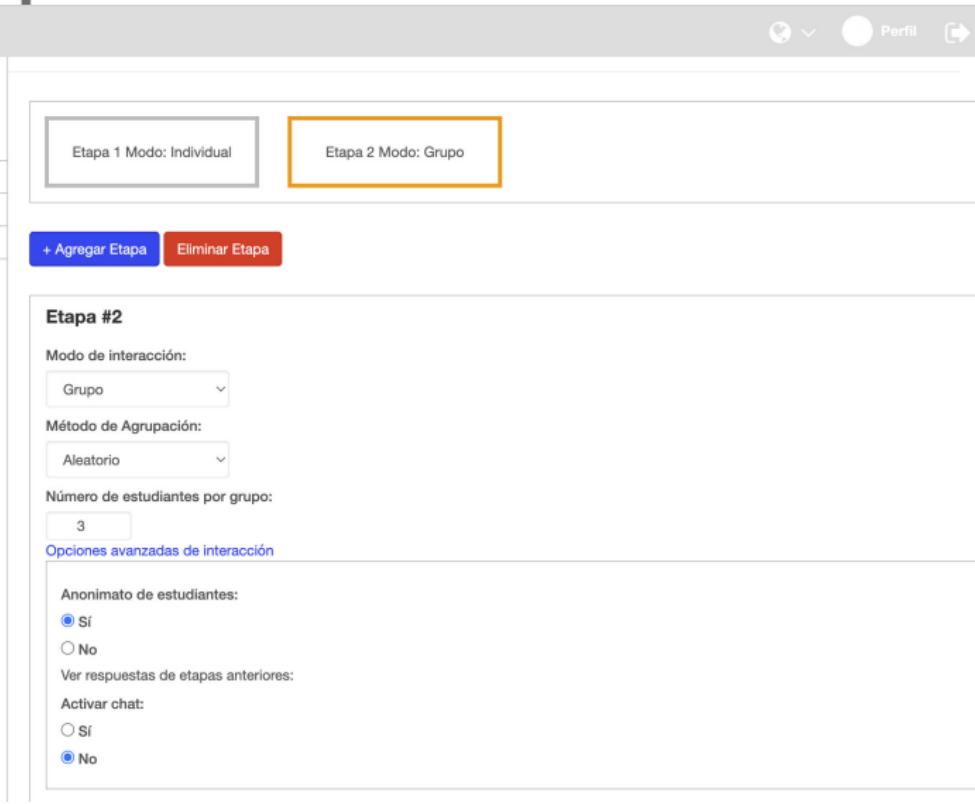
EthicApp

Inicio

Lanzar Actividad

Actividades

Diseños



The screenshot shows the EthicApp interface for creating a new activity. On the left is a sidebar with navigation links: Inicio, Lanzar Actividad, Actividades, and Diseños. The main area has a header with user profile icons and a 'Perfil' button. Below this, two boxes are shown: 'Etapa 1 Modo: Individual' (gray box) and 'Etapa 2 Modo: Grupo' (orange box). A blue button '+ Agregar Etapa' and a red button 'Eliminar Etapa' are at the bottom of this section. The next section, 'Etapa #2', is titled 'Modo de interacción:' with a dropdown menu set to 'Grupo'. It also includes 'Método de Agrupación:' (dropdown set to 'Aleatorio') and 'Número de estudiantes por grupo:' (text input field containing '3'). A link 'Opciones avanzadas de interacción' is present. At the bottom of this section are settings for 'Anonimato de estudiantes:' (radio buttons for 'Sí' and 'No', with 'Sí' selected), 'Ver respuestas de etapas anteriores:', 'Activar chat:', and another set of radio buttons for 'Sí' and 'No', with 'No' selected.

The application

The image displays two side-by-side screenshots of the EthicApp software interface, showing its functionality for managing and analyzing survey data.

Left Screenshot (Phase 3 Results):

- Header:** Configuration, Phase 1, Phase 2, Phase 3, **Phase 3**, **Finalizar**.
- Toolbar:** Editor, Users, Dashboard, Export Chat CSV, Export Answer CSV, **Phase 3**, Update, Español.
- Table Headers:** Author, Group, Item 1 (1-6), Item 2 (1-6), Item 3 (1-6).
- Data Rows:**
 - Group 17:** 17, A: 3.3 CV: 0.11, A: 1.0 CV: 0.06, A: 4.7 CV: 0.33.
 - Diego Orellana:** 18, 2 ✓ 0.10, 2 ✓ 0.3, 6 ✓ 0.3.
 - Daniela Losyza:** 18, 2 ✓ 0.8, 3 ✓ 0.2, 5 ✓ 0.1.
 - Rodrigo Balarit:** 18, 2 ✓ 0.9, 2 ✓ 0.3, 6 ✓ 0.7.
 - Group 18:** 18, A: 2.0 CV: 0.00, A: 2.3 CV: 0.25, A: 5.7 CV: 0.10.
 - Constanza Andrade Bozo:** 19, 3 0.4, 5 0.7, 5 ✓ 0.0.
 - Tomas Apablaza:** 19, 4 ✓ 0.7, 2 ✓ 0.14, 5 ✓ 0.0.
 - Martina Jacobs:** 19, 6 ✓ 0.4, 1 ✓ 0.10, 6 ✓ 0.0.
 - Group 19:** 19, A: 4.3 CV: 0.35, A: 2.7 CV: 0.75, A: 5.3 CV: 0.11.
 - Alan Moreno:** 20, 0.4, 0.3, 4 ✓ 0.3.
 - Maximiliano Ramirez:** 20, 3 0.2, 1 ✓ 0.3, 6 ✓ 0.3.
 - Javier Martinez:** 20, 3 0.6, 3 0.7, 3 ✓ 0.0.
 - Group 20:** 20, A: 3.0 CV: 0.00, A: 2.0 CV: 0.71, A: 4.3 CV: 0.36.
 - Belen Avendaño:** 21, 2 0.5, 1 ✓ 0.4, 4 ✓ 0.3.
 - Felipe Riquelme:** 21, 3 0.9, 1 0.2, 6 ✓ 0.0.
 - Nicolas Ivanovic:** 21, 3 0.10, 3 0.7, 5 ✓ 0.0.
 - Group 21:** 21, A: 2.7 CV: 0.22, A: 1.7 CV: 0.89, A: 5.0 CV: 0.26.

Right Screenshot (Item Response Detail):

- Header:** Configuration, Phase 1, Phase 2, Phase 3, **Finalizar**.
- Toolbar:** Editor, Group, Alan M., Maximiliano Ramírez, Javier Martínez, Belén A., Felipe R., Nicolás I., Fernanda L., Luis F., Bárbara G., Gustavo G., Yamila B., Francisca F., Bernardo B., Agustín A., Sebastián S., Finalizar, Cerrar, Español.
- Text:** Item3: Fue adecuado que Laura ante al error cometido en su anterior trabajo haya renunciado. - Phase3
- Scale:** Fue adecuado (1 to 6) and No fue adecuado.
- Text Response:**

C creo que a pesar de los errores cometidos no era necesario que Laura renunciara a su trabajo ya que cualquier persona puede cometer errores y en este caso no fue nada grave (según comenta el caso). A pesar de esto si ella se sentía más cómoda renunciando era la mejor opción para que desarrolle su trabajo sin problemas. (vuelvo a 4 ya que la escala se redujo)
- Scale:** Fue adecuado (1 to 6) and No fue adecuado.
- Text Response:**

A Mantengo el pensamiento que tengo en este ítem: Pienso que todo tiene su tiempo y que es bueno cumplir pequeños objetivos a corto plazo, si cometió un error, éste debería servirle como aprendizaje y a medida que pasa el tiempo lograr un progreso.
- Scale:** Fue adecuado (1 to 6) and No fue adecuado.
- Text Response:**

B Mantengo nuevamente la opinión anterior. No fue tan adecuado que Laura renunció al trabajo debido a su error pues no fue nada tan grave como para impedirle seguir trabajando ahí por lo que debería aprender de lo que hizo y seguir trabajando. Agrego que no coloco un 7 porque si ella no se sentía cómoda trabajando ahí está bien que se vaya.
- Text Response:**

B porque no fue adecuado pero capaz se siente comoda haciendo 18:00 26-09-20
- Text Response:**

C si es igual ahora como sacaron el 7 no se si es mejor dejar en 4 o en 5 18:00 26-09-20
- Text Response:**

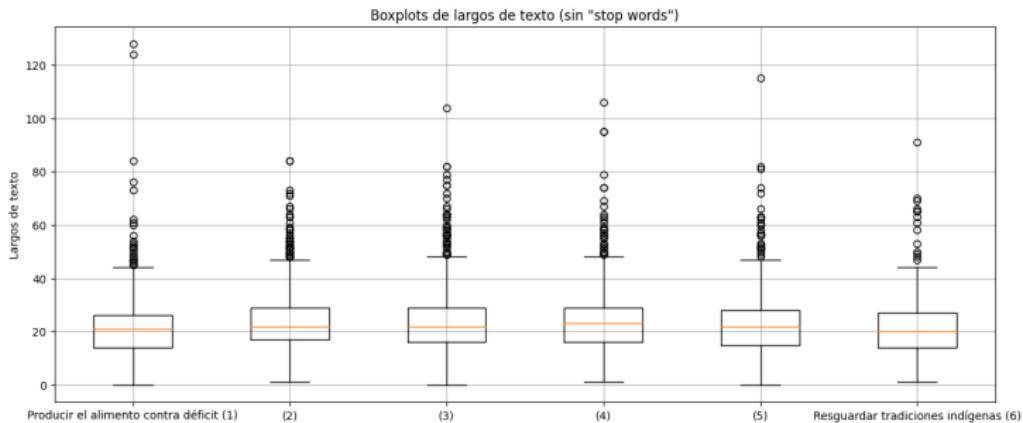
B verdad la escala cambio 18:00 26-09-20
- Text Response:**

B yo lo dejo en 5 porque envola es un error lo del 7 18:00 26-09-20
- Text Response:**

A ademas de lo anterior 18:00 26-09-20

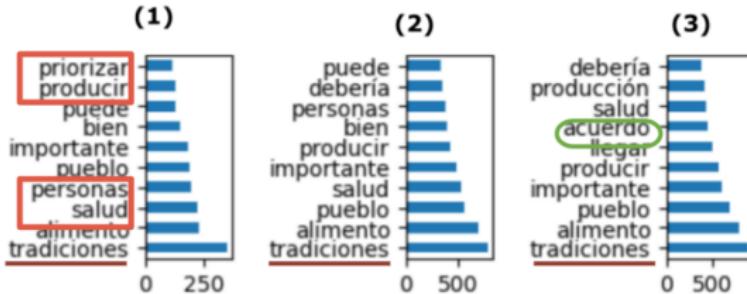
Data characteristics

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

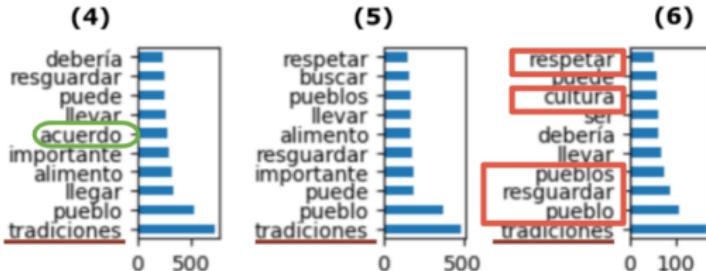


Common tokens

Produce the food against the
vitamin deficit



Protect the identity traditions



Frequency of selections

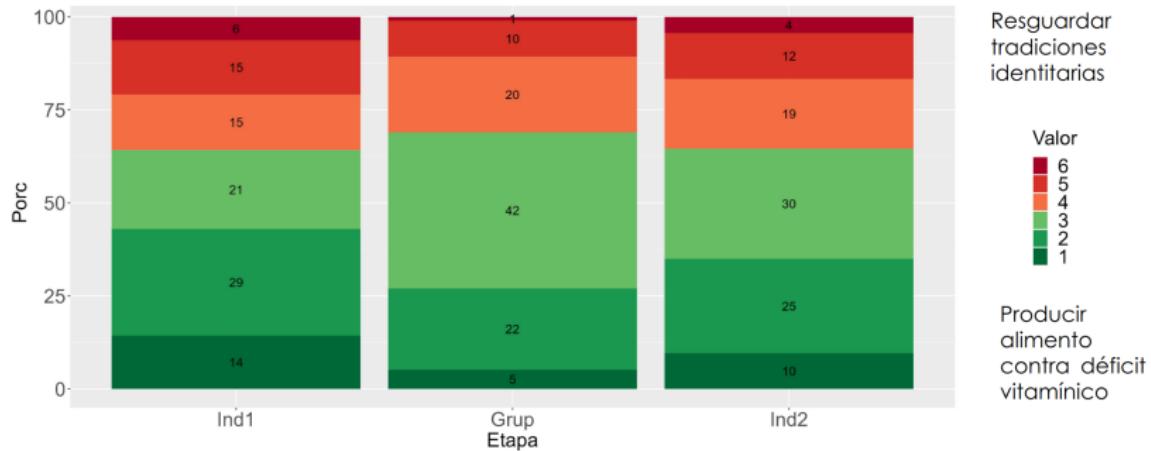


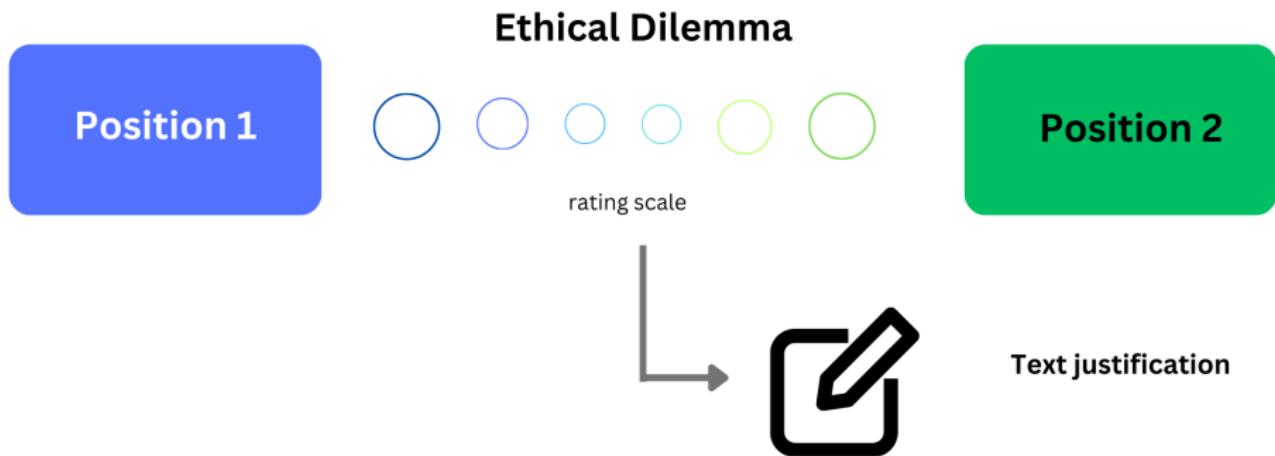
Figure: Students' choice on the Adela dilemma for a course section on each phase.

Table of Contents

- 1 Outline**
- 2 Facing the ethical dilemma: The Adela case**
- 3 Evaluating the ethical competency**
 - Ethical Formation
 - The use of dilemmas
- 4 Use of NLP in morality assessment**
- 5 EthicApp: tool and dataset**
 - The application
 - Data characteristics
- 6 Text analysis using dilemmas**
 - Predicting students answers
 - Predicting changes in position
 - Future work
- 7 References**

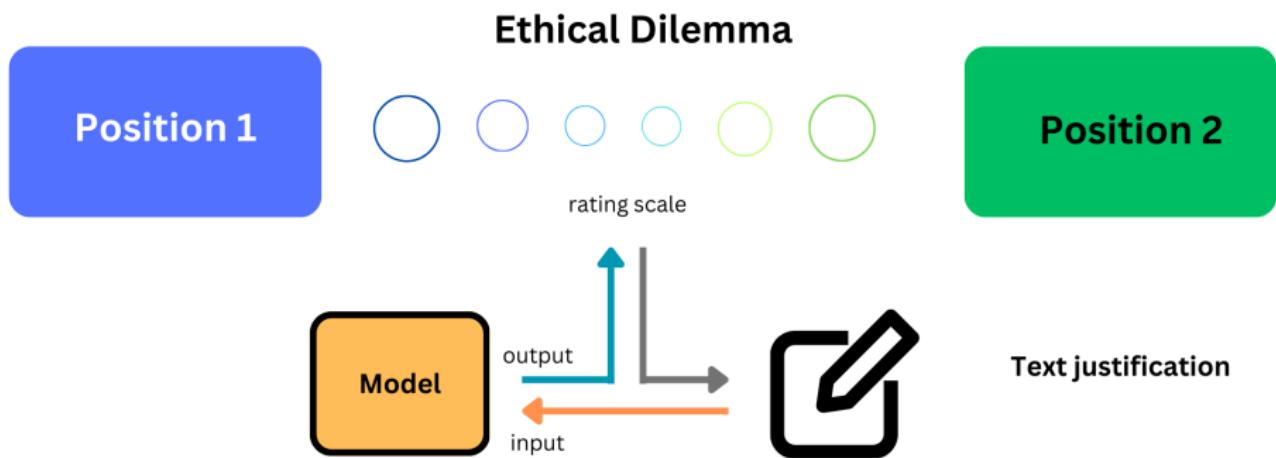
Predicting students answers

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:

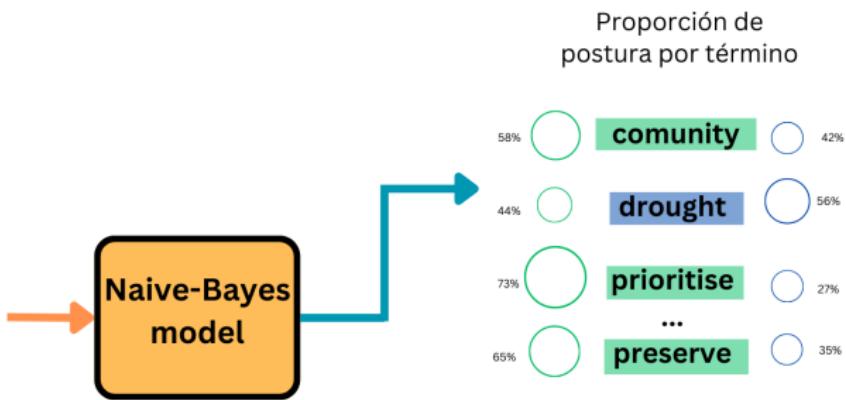
- They are interpretable,
- They allow us to perform a token-level analysis,
- They are extremely lightweight when compared to deep-learning models.

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.

Ethical Dilemma

Text justifications



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. <i>salvar</i>	0.985	6. <i>poniendo</i>	0.942	1. <i>siglos</i>	0.880	6. <i>usuario</i>	0.761
2. <i>juego</i>	0.958	7. <i>riesgo</i>	0.938	2. <i>suplementos</i>	0.863	7. <i>consentimiento</i>	0.761
3. <i>vidas</i>	0.952	8. <i>tiempos</i>	0.934	3. <i>sol</i>	0.825	8. <i>integridad</i>	0.730
4. <i>tribu</i>	0.950	9. <i>cambiarían</i>	0.934	4. <i>existen</i>	0.822	9. <i>sacar</i>	0.724
5. <i>religiosas</i>	0.948	10. <i>ayudarían</i>	0.934	5. <i>única</i>	0.770	10. <i>obtener</i>	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

Ethical Dilemma



Comment

Adela debería considerar que el sector es de escasez hídrica, por lo que un gran consumo de agua podría tener graves repercusiones en la comunidad, y a la larga podría significar que este producto no se pueda generar en unos años más. Por otro lado, se pueden buscar otras formas de obtener vitamina D para niños y ancianos. Por ello, considero que se debería priorizar preservar el recurso natural.

input

Naive-Bayes model

Probabilities for each term



58% 42%



44% 56%

Adela debería considerar que el sector es de escasez hídrica, por lo que un gran consumo de agua podría tener graves repercusiones en la comunidad, y a la larga podría significar que este producto no se pueda generar en unos años más.

Por otro lado, se pueden buscar otras formas de obtener vitamina D para niños y ancianos. Por ello, considero que se debería priorizar preservar el recurso natural.

output



65% 35%



73% 27%

Document probability

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 modifiable, 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**

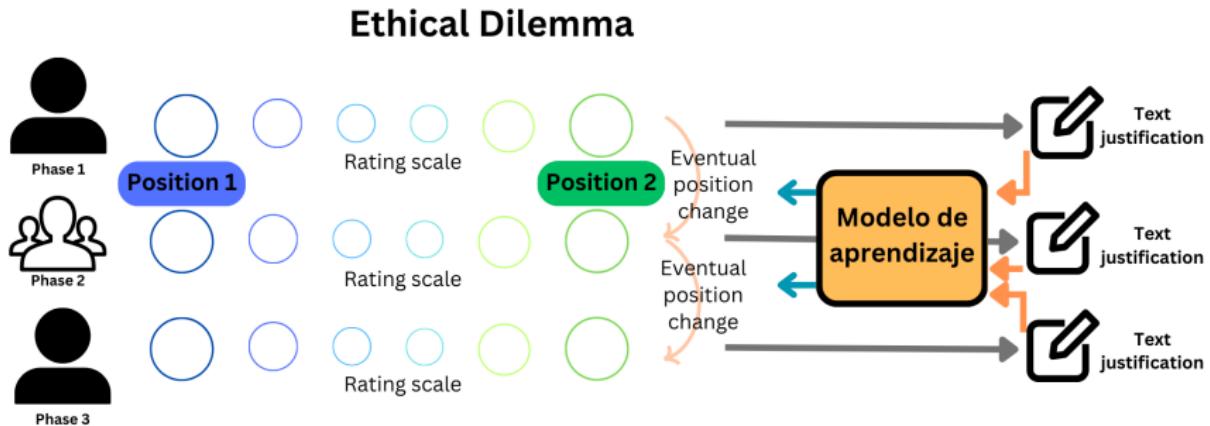
Topic Modelling

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

1 Outline

2 Facing the ethical dilemma: The Adela case

3 Evaluating the ethical competency

- Ethical Formation
- The use of dilemmas

4 Use of NLP in morality assessment

5 EthicApp: tool and dataset

- The application
- Data characteristics

6 Text analysis using dilemmas

- Predicting students answers
- Predicting changes in position
- Future work

7 References

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.

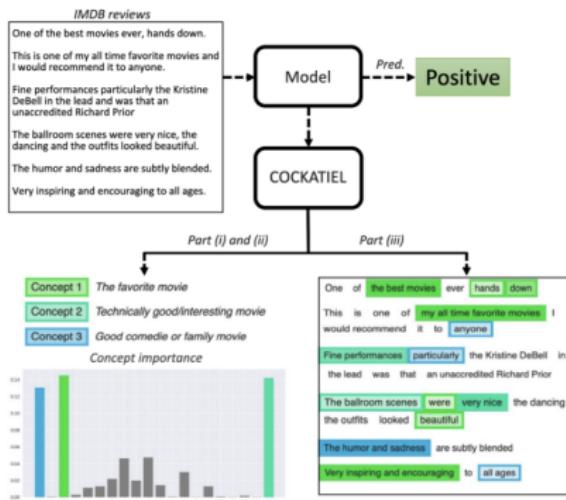


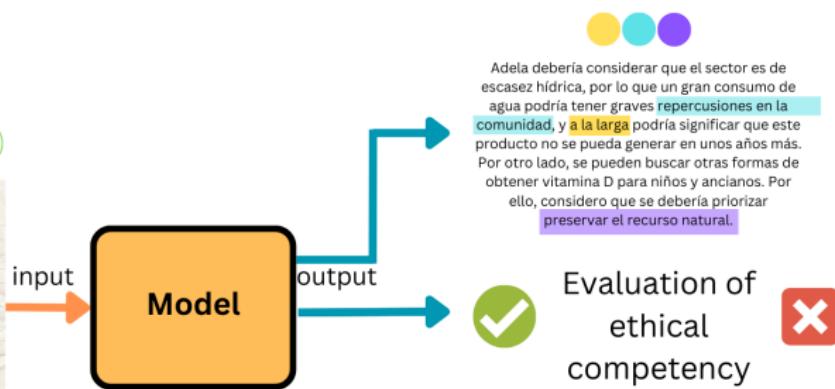
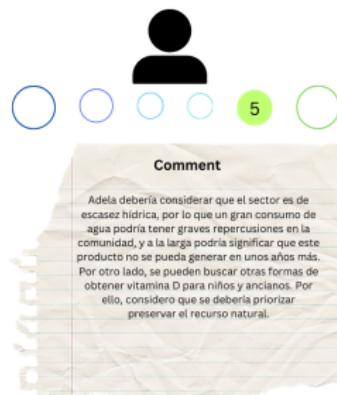
Figure 1: An illustration of COCKATIEL. Given some sentences of IMDB reviews, COCKATIEL (i) identifies concepts for prediction, (ii) ranks them, and (iii) gives the most important elements for each concept (to help us interpret the concept).

Dataset expansion

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.

Ethical Dilemma



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., Cerdá 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