Understanding the Reasoning behind Ethical Dilemmas with Interpretable Machine Learning

Camilo Carvajal Reyes
Master of Data Science
Unidad de Ética, ETHICS
Universidad de Chile
ccarvajal@dim.uchile.cl

Josefa Cerda Maureira
Unidad de Ética, ETHICS
Universidad de Chile
jozefacerda@uchile.cl

Pablo Ramírez Rivas
Unidad de Ética, ETHICS
Universidad de Chile
pabramirez@uchile.cl

Eduardo Hurtado Mila
Dpto. de Ingeniería Industrial
Unidad de Ética, ETHICS
Universidad de Chile
eduardo.hurtado@ug.uchile.cl

***Abstract*—The Ethics Unit of the Faculty of Physical and Mathematical Sciences (FCFM) of the University of Chile, has the goal of training and evaluate students in ethical competencies. One of the activities it carries out to this end is the discussion of ethical dilemmas using the tool *EthicApp*. This paper addresses the use of machine learning models to explore the textual structure of the answers and thus support the analysis that can be done by the teaching teams, which is made difficult task due to the volume of the data. Despite the potential for improvement, the methodology provides an overview of the answers and concepts used, which will contribute to the evaluation of the ethical competence of FCFM students. The results, presented using the "Adela case", allow us to conclude that students are inclined towards the principle of responsibility when faced with responsibility and that, in general, they lack the use of ethical principles in their justifications.**

***Index Terms*—ethics, natural language processing, interpretability**

I. Introduction

A. Teaching and evaluating ethics

The Ethics Unit is in charge of developing the ability of ethical commitment at the U. de Chile engineering school. The number of students in each cohort (approximately 1000) poses a challenge when it comes to analysing ethical competency. The evaluation and teaching take place in the first three semesters, during the courses from the Innovation Area. At such a stage, the desired competence they need to develop is "Reflecting 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". Indicators for the evaluation of such abilities include the principles of responsibility, respect and integrity.

B. The use of 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 (due to their consequences, the motivations involved, the adequacy to norms, etc.) 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].

II. Related works

A. Natural language processing and human morality

There are several works that seek the prediction/identification of moral elements in natural, in particular using natural language processing tools [4]–[6]. However, most of them rely on the Theory of Moral Foundations (TMF) [7], in which a set of defined morality indicators characterise the morality reasoning of a person. Works such as [6] use language processing techniques (including *word embeddings*, DFM occurrence counting [8] and BERT [9]) to predict users' stance towards each of the moral dimensions posed in the TMF [7], using text written by them on social media. We follow this line of work regarding the machine learning algorithms, but differ in the use of morality: the Ethics Unit assumes that ethical thinking can be taught instead of being predetermined. Moreover, as we will see in section II-B, our task is a particular one in which the students are requested to show their ethical reasoning directly, which makes it different from most existing works in the literature.

B. The EthicApp application

EthicApp consists of an application that allows for the implementation and monitoring of ethical dilemmas [10] . We show the Adela case in order to contextualise the activities carried out by the students: *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 in order to conserve vitamin D during transport to other areas, the team decided to freeze-dry the fruit and add preservatives. Adela listens to the Diaguita community (despite not being legally obliged to do so), which states that the procedure goes against the communities' traditional practices, which are a fundamental part of their identity.* Adela faces the dilemma of prioritising the populations' health by producing the vitamin or to respect the identity traditions of the Diaguita community. Students need to choose between the two options, giving a score on a scale from 1 to 6.

III. Methodology

We propose the use of models for predicting the score on the rating scale between the two positions given by the student, using the text justification, as shown in Fig. 1. 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 uses deep learning techniques in a follow-up work from their first NLP-based assistance tool [11].

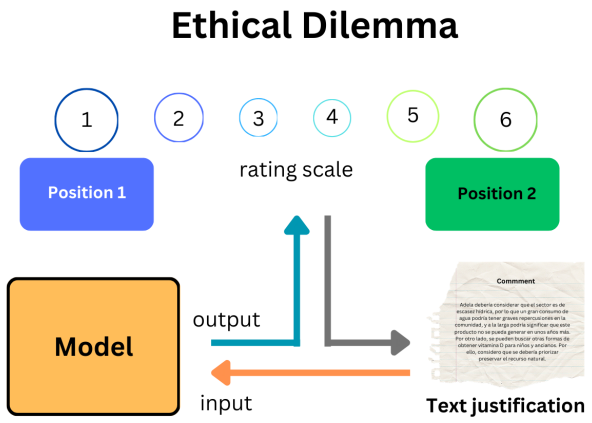

Fig. 1. Proposed use of ML methods for predicting the position regarding the dilemma.

During the first stage of the project, we give priority to simple methods based on a bag-of-words vectorisation, i.e., using the occurrence of each word in the vocabulary as a feature vector. We then utilise methods including Naive-Bayes, linear models and ensemble methods (random forest and boosting). Despite their use having decreased at the expense of large language models (LLMs), their use is convenient in this context since they are interpretable, allow us to perform a token-level analysis and they are extremely lightweight when compared to deep-learning models. The models naturally assign a score to each word (probability, weight or feature importance) with which it is possible to identify those elements that contribute more to the final result of the algorithm, obtaining the interpretative capacity. Tests are carried out by separating the six positions as independent cases, as well as grouping the options in two.

IV. Results

A. Method comparisons

The following table shows the general results for the interpretable models that were tested. Both classifier metrics were used (summarised using $f1$) and regression ones (using the $R2$ score and the mean absolute error) since the task can be regarded both ways. We also include the simplified task of predicting the general position taken by grouping together selections 1 to 3 and 4 to 6 (first column).

Model	Binary f1	f1	R2	MAE
Naive-Bayes	**0.79**	**0.43**	**0.17**	**0.78**
Linear	**0.82**	**0.44**	**0.38**	**0.77**
Random Forest	**0.77**	**0.38**	**0.34**	**0.82**
Gradient Boosting	**0.77**	**0.39**	**0.35**	**0.81**

B. Visualisation of token probabilities/logits

We provide a prototype of a visualisation tool, depicted in Fig. 2. 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. It is intended to complement the inspection that academic teams need to do of students' answers.

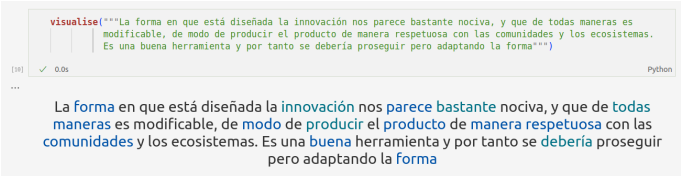

Fig. 2. Example of visualisation of feature importance within a given justification to the dilemma.

C. Qualitative analysis of features

After the analysis of concepts for which a high probability was identified for each of the positions using the Naive-Bayes model as a reference, the presence of these concepts in the text correlates strongly with the positions in question. The presence of these concepts in the text correlates strongly with the positions in question, which highlights the most recurrent semantic elements when writing the justifications. A similar analysis can be carried out with the linear models. The table shows that for the position of producing the food, words such as "save", "game", "lives", "tribe", "religious" and "risk" are more likely to be used in the justification; these words could show that those who choose this position consider that saving lives or the lives at stake are more important. They also refer to indigenous peoples as "tribe" and to their traditions as "religious". On the other hand, the words that stand out from those most likely to be used for the stance of safeguarding identity traditions are "supplements", "sun" and "obtain"; this could be related to those who choose this option pointing to alternatives to vitamin deficiency via supplements or obtaining it through the sun. This analysis of representative tokens and students' preferences shows that in general, the principle of responsibility prevails over the principle of respect

V. Further work

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

[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] 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

[5] Xie, J. Y., Hirst, G., & Xu, Y. (2020). Contextualized moral inference (arXiv:2008.10762). arXiv.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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