Classifying (Un)Ethical Workplace Dilemmas: A Survey of Machine Learning Approaches

Breanna Nguyen
LING 227: Language and Computation | Professor Tom McCoy
May 6, 2024

Abstract

A wide range of approaches can be applied to the task of making ethical decisions. One could use a utilitarian approach, only considering relative benefits and harms. One could use a deontological approach, only considering the moral principles that govern their decisions. One could simply act on intuition alone. Each approach uses different information and processes to come to a decision. From an computational perspective, ethical decision-making can be implemented as algorithms that take in data and output a decision. In this project, I implemented three computational models to classify ethical dilemmas in the workplace: naive Bayes, logistic regression, and classification using large language models. My goal was to investigate which models and types of data produce outputs that reflect human decisions based on a corpus of ethical dilemmas in the workplace. The results of evaluating each model's performance suggest that logistic regression is the most appropriate classification method given the nature of the data. I concluded by discussing the overall performance of all the models and suggested future work in this domain.

1 Introduction

Ethical decisions are an important aspect of the human experience. In many situations, we often appraise the actions of others and ourselves to come to a decision about who is in the right or wrong. How we come to those decisions is influenced by a myriad of factors, including our personal and moral values, and the specificities of the situation. Those decisions are also often hard to characterize, as they can be subjective and context-dependent.

With the recent emergence of machine learning techniques comes the question of if they can be applied to ethical decisions. The field of computational ethics aims to develop algorithms that can make ethical decisions (Awad et al., 2022). There are many aspects of this goal that remain unclear including which values the decisions should be based on, which algorithms or models are appropriate for the task, and the ethical implications of using a machine to make ethical decisions.

In the project, I aimed to answer the following research question: Which method is most effective at classifying ethical dilemmas in the workplace and what are the implications of these findings? I implemeted three models to answer this question: naive Bayes, logistic regression, and classification using large language models (LLMs). I evaluate the performance of each model and end with a discussion of the results and points for further investigation.

Previous work in computational ethics includes GenEth, an ethical dilemma analyzer that uses inductive logic programming to codify ethical principles, The Moral Machine, which explores the moral dilemmas of autonomous vehicles, and a fuzzy-cognitive maps approach to decision-making in medical ethics (Anderson & Anderson, 2018; Awad et al., 2018; Hein et al., 2022). This project differs from the current work in computation ethics because it takes a more observational, as opposed to experimental, approach to data collection. Data was collected from posts about workplace dilemmas on the internet.

Attempting to combine computation and ethics raises the concern of if machines should be involved in ethical decisions *at all*. Machines often evoke a false sense of objectivity, which can be used to justify harmful decisions (Blackwell, 2019). The decisions of machines are ultimately based on the humans that created them and are therefore subjective. Furthermore, implementing machine learning models that make ethical decisions requires a process of converting the ethical dilemmas into quantitative features or expressing them through natural language. Both of these practices reduce the human experience into fewer dimensions and therefore do not fully or accurately reflect the dilemma. These issues remain unresolved and provide an opportunity for further investigation. For the purpose of this project, I accept that the aforementioned issues persist and proceed with them in mind.

2 Data Description

2.1 Collection and Cleaning

Data were collected from a Reddit-based forum called "AmITheAsshole" (AITA), where users post about ethical dilemmas with the expectation that other users will tell them if they're in the right or in the wrong. In each post, the original poster (OP) briefly describes the dilemma they faced and relevant contextual information. Other users comment NTA (Not The Asshole) or YTA (You're The Asshole) based on their assessment of the situation. The ground truth evaluation ('label' in Table 1) is measured as the ratio of NTA votes to all votes. Ratios at or above 0.500 are labeled NTA and ratios at or below 0.499 are labeled YTA.

Posts were filtered for the keywords "work", "job", "coworker", "colleague", "boss", "manager", "supervisor", "employ", "employee", and "employer". They were then sorted by popularity in descending order and screened for eligibility. Posts were included if the content was about OP and another person (the target) and excluded if they were about someone who does not work with OP. if they were hypothetical situations, or OP was acting against more than one person.

I defined the variables based on what was usually available in posts on this forum about the workplace. For example, users usually included their gender age, the target's gender and age, how long they have had that job, what action they took, and the impact of that action. A full list of variables collected is available in Table 1. All posts were manually codified.

2.2 Variables

Table 1: Description of Variables in the Dataset

Variable	Description
title	Title of the post
content	Content of the post
op_gender	OP's gender
op_age	OP's age
target_gender	Target's gender
target_age	Target's age
length	Length of time at that job
experience	Amount of experience (relative to target)
relationship	Job position (relative to target)
industry	Industry of the workplace
condition	Presence of existing condition
action	Action taken by OP
intention	Intention behind the action
impact	Impact of the action
good_standing	Is OP in good standing?
perspective	Does OP consider another perspective?
label	Verdict

Many of the variables are clear given their short description above, but some require further explanation. Condition indicates if a pre-existing workplace condition existed. This can include an established rule, an existing agreement, or approved time off. Action, intention, and impact are categorical variables with between 8 and 10 levels. Good standing indicates whether OP mentions that they are in good standing at their job. Perspective indicates whether OP considers a perspective other than their own about the situation. Full lists of the action, intention, and impact variable levels are available in the supplemental information.

3 Methodology

3.1 Data Preprocessing

The categorical variables were converted to numerical form via one-hot encoding. Missing values were filled in with the respective mode for categorical variables and mean for numerical variables. The entire data set was scaled by dividing each value by the standard deviation of the respective variable.

The data were converted from csv (comma separated values) form to txt (text) and xml (extensible markup language) to be compatible with the machine learning methods of this project. The txt (simple) files only contained categorical variables and the label, and were split by an 80/20 ratio of training and test rows. The xml (full) files contained the complete set of variables, including the tokenized full text of the post, and the label, and were split by an 70/20/10 ratio of training, test, and development rows. The full files also contained the following extended features: sentiment (using the Opinion Lexicon from Liu et al.,

2005) and presence of curse words. The data were split into the simple and full files so I could evaluate the performance of each model on different subsets and types of data.

3.2 Naive Bayes

The first method I implemented was a naive Bayes classification model. This model utilizes Bayes' Theorem to output a class, c, given a document, d. To compute the class, I can use two other elements that are a bit easier to compute: $P(d|c)$ and $P(c)$. These terms constitute the numerator of the application of Bayes' theorem to this task. The denominator is disregarded because it is the same in every case. I get $P(c)$ by calculating the proportion of each label in the training set, which was already manually labeled. I get $P(d|c)$ finding the proportion of d found in class c. A smoothing parameter (lambda) can be added to avoid outputting a probability of zero in the case of unseen data.

For the simple set, d consists of only categorical features. In this case, the naive Bayes classification model outputs a label of NTA or YTA given the set of categorical features of a post.

The naive Bayes model performed moderately well at classifying this data set. It achieved an accuracy of 71.43%, which is much higher than at-chance guessing based on the proportion of NTA's in the training set, 60.71%. The accuracy remains unchanged with the addition of smoothing parameters 1 and 2, but drops to 66.67% with smoothing parameters of 3 and above.

For full files, d consists of a bag-of-words representation of the post, as well as all of the other features. In this case, the naive Bayes classification task outputs a label of NTA or YTA given the tokenized set of words in a post and the categorical features of a post.

The naive Bayes model also performed moderately well at classifying this data set. It achieved an accuracy of 72.73%, which is much higher than at-chance guessing based on the proportion of NTA's in the training set, 62.16%. The accuracy remains unchanged with the omitting of extended features. Out of curiosity, I ran the model with only the ethical variables, action, intention, and impact, as features and achieved the same accuracy.

A table with the precision, recall, and F1-score values from the version of the model with the best accuracy is available below.

Table 2: Naive Bayes Performance

Label	Precision	Recall	F1
NTA	62.50	100.00	76.92
YTA	100.00	50.00	66.67
Overall Accuracy:	72.73		

3.3 Logistic Regression

The second method I implemented was a logistic regression classification model with the help of the OpenNLP MaxEnt Framework (Baldridge et al., 2011). This model differs from the naive Bayes model in that it extended beyond just using probabilities. This model utilizes the softmax function to convert the linear combination of input features to a probability distribution over the different classes in the data set. This probability represents the likelihood that the given input belongs to a certain class. The class with the highest probability is the one that the model assigns to the input. A smoothing parameter (sigma) can be added to change how extreme the parameters of the model are.

For the simple set, the logistic regression model outputs a label of NTA or YTA given the set of categorical features of a post.

The logistic regression model performed moderately well at classifying this data set. It achieved an accuracy of 71.43%, which is much higher than at-chance guessing based on the proportion of NTA's in the

training set, 60.71%. The table below shows the change in accuracy as I tried different smoothing parameters. The highest accuracy achieved with the simple set was 76.19% with a smoothing parameter of 0.1.

Table 3: Logitic Regression Performance by Sigma Value (Simple Data)

Sigma	.01	0.1	0.5	1.0	5.0
Accuracy	66.67	76.19	71.43	71.43	71.43

For the full set, the logistic regression model outputs a label of NTA or YTA given the tokenized set of words in a post and the categorical features of a post.

The logistic regression model also performed moderately well at classifying this data set. It achieved an accuracy of 72.73%, which is much higher than at-chance guessing based on the proportion of NTA's in the training set, 62.16%. The table below shows the change in accuracy as I tried different smoothing parameters. The highest accuracy achieved with the full set was 72.73% with a smoothing parameter of 1. This level accuracy remained unchanged with the omitting of extended features.

Table 4: Logitic Regression Performance by Sigma Value (Full Data)

Sigma	.01	0.1	0.5	1.0	5.0
Accuracy	54.55	63.64	63.64	72.73	72.73

A table with the precision, recall, and F1-score values from the version of the model with the best accuracy is available below.

Table 5: Logistic Regression Performance

Label	Precision	Recall	F1
NTA	75.00	92.31	82.76
YTA	80.00	50.00	61.54
Overall Accuracy:	76.19		

3.4 ChatGPT vs. Gemini

The third method I implemented was LLM-enabled text classification via publicly available online Large Language Models (LLMs), ChatGPT and Gemini. Contrary to the previous two methods I implemented, this method only utilized the text content of the post with no additional features. Additionally, this method did not require any training data. These LLMs were fined-tuned on a large corpus of text to execute next-word prediction tasks. They are based on the transformer architecture, which is a type of artificial neural network that is well-suited processing sequences of data. I employed ChatGPT and Gemini to predict the label of right or wrong (which was later mapped on to NTA and YTA) given the text content of the post. However, these LLMs have the tendency to be overly verbose in their responses which was remedied by prompt engineering on my part. For each test post, I pasted the text content followed by this sentence: "Complete the following sentence with either"right" or "wrong": In terms of ethics, the narrator of this story is in the". I then re-coded the output to NTA or YTA based on the response.

Two tables with the precision, recall, and F1-score values from ChatGPT and Gemini are available below.

Table 6: ChatGPT Performance

Label	Precision	Recall	F1
NTA	88.89	53.33	66.67
YTA	41.67	83.33	55.56
Overall Accuracy:	61.90		

Table 7: Gemini Performance

Label	Precision	Recall	F1
NTA	81.82	60.00	69.23
YTA	40.00	66.67	50.00
Overall Accuracy:	61.90		

3.5 Supervised Text Classification with GPT-2 & Classification with BERT Embeddings

The fourth and fifth methods I attempted were a supervised text classification model using GPT-2 and a classification model using BERT embeddings. Both of these models were fine-tuned on a large corpus of text but GPT-2 is tuned for text generation and BERT was tuned to generate embeddings for text classification. Training both of these models on example data (unrelated to this project) was successful, but I was unable to train it on my own data as the content of the posts was much longer than what each model could handle.

4 Results

In this project, I implemented three models to classify ethical dilemmas in the workplace: naive Bayes, logistic regression, and LLM-enabled text classification via ChatGPT and Gemini. I also attempted to implement supervised text classification using GPT-2 and classification using BERT embeddings without success due to resource contraints within the models. Of the models I successfully implemented, the logistic regression model, trained and evaluated on the simple data set, performed the best with an overall accuracy of 76.19%. In order from most to least accurate, the models performed as follows: logistic regression (simple data), naive Bayes (full data), logistic regression (full data), naive Bayes (simple data), ChatGPT, and Gemini.

Precision, recall, and F1-score are metrics to help address the imbalance of classes in the data. Precision is the ratio of correctly predicted observations to the total predicted observations of a class. Recall is the ratio of correctly predicted observations to the total actual observations of a class. F1-score is the harmonic mean of precision and recall. Knowing these values allows me to evaluate the tradeoffs between each model's performance. In general, all models seem to have a higher F1-score and precision for NTA observations than YTA observations This makes sense given the higher proportion of NTA observations in the data set.

Recall is also higher in NTA observations than YTA observation for the naive Bayes and logistic regression models, but the opposite is true for classification with ChatGPT and Gemini. This means that ChatGPT and Gemini were worse as classifying NTA than they were at classifying YTA. This could be due to the lack of training for these models or the quality of prompt engineering.

Of all the models, logistic regression trained on the simple data performed still performed the best when we consider the tradeoffs between precision, recall, and F1-score.

5 Discussion

The results of this project suggest that logistic regression is the most appropriate model for classifying ethical dilemmas in the workplace given the specific structure of my datasets and a comparison to naive Bayes and online LLMs. This is a striking result because not only is logistic regression a much simpler model than what is utilized by ChatGPT and Gemini, but the best version was also trained on a very simple version of the data. In the context of this project, this means that increased complexity in the model and data do not correlate with an increase in performance.

Additionally, logistic regression model also yieled the same overall accuracy as a K-Nearest Neighbors algorithm I implemented previously but did not include in this report. K-Nearest Neighbors is a non-parametric method that classifies data points based on the majority class of its k-nearest neighbors using Euclidean distance. This suggests that both parametric and non-parametric methods are appropriate for this task.

There are a number of limitations to consider and opportunities for further exploration as a result of this project. The first limitation is that the data set is small, which could have lead to inadequate training. Furthermore, the data set was collected and annotated without any cross-validation so the features are biased towards solely my evaluation. The second limitation is that the naive Bayes and logistic regression models assume some extent of independence in the data, which is not necessarily true. This can lead to misleading conclusions about how these models perform. The third limitation is that the data could have appeared in the training set for the LLMs, which could have biased their output. Lastly, the lack of resources to train GPT-2 and BERT on my data is a limitation of this project.

Further work is required to address these limitations. First, the data set should be expanded to include more posts and there should be additional people to cross-validate the annotations. Second, the models should be re-implemented with a more complex structure that accounts for the interdependence of the data. Third, the LLMs should be introduced to data that has not appeared in their training set to get a true sense of their performance. Lastly, I could re-attempt to implement GPT-2 and BERT with greater resources such as high performance computing clusters and increased memory.

In conclusion, this project surveyed a variety of methods for classifying ethical dilemmas in the workplace based on the content of posts from a Reddit-based forum. The data sets consisted of the text of the post and hand-annotated categorical features. Given the structure of the data, the logistic regression model trained on the data set with only categorical features performed the best when compared to classifcation using naive Bayes and LLMs. Future work in this domain should try to improve the performance of these models as well as survey additional methods for classifying ethical dilemmas. Still remaining are the ethical and societal implications of such algorithms.

References

Anderson, M., & Anderson, S. L. (2018). GenEth: A general ethical dilemma analyzer. Paladyn, Journal of Behavioral Robotics, 9(1), 337–357. https://doi.org/10.1515/pjbr-2018-0024

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A. C., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., & Tenenbaum, J. B. (2022). Computational ethics. Trends in Cognitive Sciences, 26(5), 388–405. https://doi.org/10.1016/j.tics.2022.02.009

Baldridge, J., Morton, T., & Baldridge, J. (2011). The OpenNLP Maximum Entropy Package. https://maxent.sourceforge.net/about.html

Blackwell, A. F. (2019). Objective functions: (In)humanity and inequity in artificial intelligence. HAU: Journal of Ethnographic Theory, 9(1), 137–146. https://doi.org/10.1086/703871

Hein, A., Meier, L. J., Buyx, A. M., & Diepold, K. (2022). A Fuzzy-Cognitive-Maps Approach to Decision-Making in Medical Ethics. 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–8. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882615

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: Analyzing and comparing opinions on the Web. Proceedings of the 14th International Conference on World Wide Web - WWW '05, 342. https://doi.org/10.1145/1060745.1060797

Code Availability

https://github.com/breannaknguyen/LING227_proj

Code was adapted from LING 227 homework 3 and LLM discussion section.

Supplemental Information

Additional Variable Descriptions

Table 8: Description of 'Action' Levels

Level	Description
Telling the truth	OP honestly conveys information without deception or distortion
Lying or misrepresenting information	OP provides false or deceptive information
Cheating or violating company rules	OP uses dishonest, unethical, or rule-breaking methods to gain an unfair advantage
Stealing or misusing company resources	OP uses company resources, such as supplies or time, without authorization or proper purpose
Verbally insulting or disrespecting	OP uses offensive or demeaning language, or condescending tones with the target
Ignoring or neglecting duties or people	OP fails to fulfill their responsibilities, or provide attention to something that is expected of them
Violating professional boundaries	OP engages in behavior that transcends appropriate boundaries between personal and professional relationships
Making unreasonable or declining reasonable requests	OP asks something of or refuses a request from the target
Disciplinary/professional action	OP appeals to administrative power to act against the target

Table 9: Description of 'Intention' Levels

Level	Description
Self advancement or interest	OP acted with a positive outcome for themselves or only their own interest in mind
Helping others	OP wanted to benefit others in the workplace
Seeking fairness or justice	OP wanted equitable treatment for themselves or others
Personal vendetta	OP was acting on pre-existing dislike for the target
Asserting boundaries	OP wanted to uphold their personal or professional limits
Avoiding conflict	OP wanted to minimize tension or disagreement in the workplace
Retaliation	OP was responding to an injustice in a congruent manner
Reinforcing authority	OP used their power or position to act against the target
Conforming to norms	OP wanted to uphold social or workplace norms
Following rules	OP wanted to uphold explicit rules of their workplace

Table 10: Description of 'Impact' Levels

Level	Description
Strained relationships	There are tense or uncomfortable relationships amongst people in the workplace
Productivity decrease	There is a decline in efficiency in the workplace
Damage to reputation	There is a negative impact on the workplace's or another person's standing
Emotional distress	There are feelings of sadness, discomfort, or anxiety in one or more persons in the workplace
Emotional distress	The workplace or an individual suffers negative financial impacts
Career consequences	An individual suffers negative impacts on their career such as a demotion or disciplinary infraction
Safety concerns	There is a risk to the physical wellbeing of persons in the workplace
Disruption to workplace dynamic	There is a negative disturbance to the usual workplace environment such as a changed rule that affects many people or a change in the way the people in the workplace interact

Terminal Commands

Train and evaluate naive Bayes (simple)

```
./naivebayes.py -t data/simple/train.txt -p data/simple/test.txt | ./score.py -g
data/simple/test.txt
```

Train and evaluate naive Bayes with smoothing parameter (simple)

```
./naivebayes.py -t data/simple/train.txt -p data/simple/test.txt -l 1 | ./score.py -g
```

Train logistic regression (simple)

```
classify train data/simple/train.txt out/aita_log_model.txt
```

Evaluate logistic regression (simple)

```
classify apply out/aita_log_model.txt data/simple/test.txt | ./score.py -g
data/simple/test.txt
```

Evaluate logistic regression with smoothing parameter (simple)

```
classify train -sigma .1 data/simple/train.txt out/aita_log_model.txt classify apply
out/aita_log_model.txt data/simple/test.txt | ./score.py -g data/simple/test.txt
```

Train and evaluate naive Bayes (full, without extended features)

```
./post_sentiment.py -t data/full/train.xml -e data/full/dev.xml -m nb
```

Train and evaluate naive Bayes (full, with extended features)

```
./post_sentiment.py -t data/full/train.xml -e data/full/dev.xml -m nb -x
```

Train and evaluate logistic regression (full, without extended features)

```
./post_sentiment.py -t data/full/train.xml -e data/full/dev.xml
```

Train and evaluate logistic regression (full, with extended features)

```
./post_sentiment.py -t data/full/train.xml -e data/full/dev.xml -x
```

Train and evaluate logistic regression with smoothing parameter (full, with extended features)

```
./post_sentiment.py -t data/full/train.xml -e data/full/dev.xml -x -s 1
```