HMM, Is This Ethical?
Predicting the Ethics of Reddit Life Protips

Madison Coots
Management Science
and Engineering
Stanford University
mcoots@stanford.edu

Peter Lu
Computer Science
Stanford University
peterlu6@stanford.edu

Lucy Wang
Computer Science
Stanford University
wanglucy@stanford.edu

Abstract

We propose an augmented hidden Markov model classifier for the purpose of ethics classification in text posts from the social media site Reddit, in the subreddits Unethical Life Protips and Life Pro Tips. Our classifier achieves an F1 score of 0.82, on par with human performance. Through examination of performance of the hidden Markov model with ambiguous data points, we demonstrate the value of intelligently falling back to different model architectures to maximize performance while maintaining interpretability.

1. Introduction

Given the burgeoning popularity of online discussion forums and social media, there has been considerable research aimed at building models to discern the sentiment of online text data including product reviews, articles, and social media posts. Other areas of computational research have yielded forms of analysis that extract overall content and main topics of texts. While there has been substantial research in both of these areas, the intersection has been largely unexplored. That is to say, there has been limited research exploring the complex interactions of tone, sentiment and content in natural language and the contributions of these factors to a texts ethical valence. This project aims to explore the intersection of these textual aspects through delving into the construction of an ethics classifier for natural language. This project's objective will pose a unique challenge in not only necessitating the extraction and capture of these characteristics from a text, but also in effectively leveraging them for an accurate prediction.

Reddit, a site for web content rating and discussion, contains a plethora of natural language text implicitly tagged as either ethical or unethical. Specifically, it contains two topic pages (also known as subreddits) pertaining to two different types of protips: Unethical Life Protips (ULPT) and Life Protips (LPT). While the ULPT subreddit largely contains tips of questionable legality or tips that may improve your life at the expense of others, LPT aggregates suggestions that are purely well-intentioned and helpful. This diverse and rich corpus of text will serve as the data used to construct the ethics classifier.

2. Literature Review

Attempting to computationally quantify the notion of ethics is an inherently challenging problem due to the high amount of subjectivity surrounding discussions on ethics. The two main approaches in the field of computational ethics are: 1) to emphasize the study and understanding of human cognitive processes in moral decision making, and 2) to to attempt to formulate a general moral framework that can be matched by computational ethics systems, even if their internal workings are far from human models of thinking. [1] Because of a relatively poor understanding of the neuroscience behind ethical decision making, it is more valuable at this point to formulate a realistic ethical framework, and attempt to achieve good performance on that using existing well-understood machine learning models (particularly those suited for textual analysis). We are using the crowdsourced ethical framework defined by a broad set of internet users (Reddit users) and attempting to match this human classification using machine learning logic.

Hidden Markov models are a powerful tool in natural language processing applications, and have been used widely in text classification. One recently implemented structure of HMMs used for binary classification involved separate HMMs trained on the positive and negative sets of data [2]. The HMMs are fed feature vectors derived from NLP approaches such as TF-IDF and ECE (expected cross entropy). Another powerful HMM structure was the usage of large ensembles [3], where HMM models are iteratively created and trained to minimize covariance and maximize predictive power, achieving extremely high accuracies eclipsing human performance.

One recently developed feature representation of text are GloVe vectors, which uses a log-bilinear regression model to dynamically generate feature vectors based on a large corpus of text. GloVe representations have been trained on several large datasets that are publicly available, including corpuses from Wikipedia, Gigaword, Twitter, and Common Crawl [4].

3. Task Definition

Our problem is one of constructing a binary ethics classifier that outputs a label: ethical or unethical, for a given natural language string. One of the considerations to be kept in mind in construction of the model is preserving the ordering of the words in the input original string. This will ensure that our model is more robust with regard to how a word is used in a sentence, rather than just focusing on simply its presence or absence from an input string.

4. Data

We used BigQuery to query posts from both the LPT (ethical) and ULPT (unethical) subreddits. Each data record represents a submitted post that is correspondingly tagged as either ethical (1) or unethical (0). Additionally, Reddit moderators regularly remove posts that are misclassified: ethical posts are disallowed in the unethical subreddit, and illegal posts are disallowed in the ethical subreddit. This allows us to use the queried data with a higher level of confidence. In total we have downloaded approximately 40,000 posts from January 2017 to August 2018. Half of the data records are labeled as ethical, and half are labeled unethical. We partitioned the data into three sets: test (10%), dev (10%), and training (80%).

Protip	**Label**
Refill your empty bottle of body wash at your gym's showers. It's Usually behind a curtain, so there's no need to be sneaky. Complementary body wash with gym membership.	0
If you have squirrels on your bird feeder, smear the supporting pole with Vaseline.	1

Table 1: Sample data records from both LPT and ULPT

4.1. Data Processing

Punctuation was separated from the words, and considered its own token. This was done to avoid having sparse tokens such as ”afford?” which would be difficult to learn.

For example, the phrase ”See someone advertising a second hand car that you can't quite afford? Key the car to bring the price down.” is transformed into ”see someone advertising a second hand car that you cant quite afford ? key the car to bring the price down .”

Each word in a given protip string, was converted into a GloVe vector. We tested GloVe vectors generated from three different corpuses of text: Twitter, Wikipedia + Gigaword, and Common Crawl. We also test vectors of varying dimensionalties ranging from $\mathbb{R}^{25}$ to $\mathbb{R}^{300}$. [4]

Testing of various GloVe vector corpuses and dimensionalities demonstrates that a 100D vector representation based on the Twitter corpus is best.

corpus	dimensionality	f1
Twitter	25	0.63
Twitter	50	0.74
Twitter	**100**	**0.77**
Twitter	200	0.75
Common	300	0.76
Wikipedia	50	0.73
Wikipedia	100	0.74
Wikipedia	200	0.75
Wikipedia	300	0.73

Table 2: GloVe Representation Testing

F1 scores were produced from a preliminary untuned HMM + ensemble backoff model with no bias.

As expected, the Twitter corpus's understanding of social media content enabled it to achieve the best performance.

5. Background on Hidden Markov Models

Hidden Markov Models are used to represent probability distributions of observing a sequence of outcomes. Additionally, HMMs are based on two key assumptions. Firstly, the observation X_i at time i is generated by a hidden process or state S_i. Secondly, HMMs abide by the Markov property: the output of S_i, given the output of S_{i-1}, is independent of all states prior to S_{i-1}. Therefore, using these properties, we are able to compute the probability of observing a sequence of T total outputs **X** given our ser of hidden states **S**:

$$P(\mathbf{X}) = \sum_{S} \prod_{i=1}^{T} P(X_i|S_i) \cdot P(\mathbf{S}) \tag{1}$$

6. Methods

In an effort to classify Reddit Life Protips based on their ethical qualities, we are modeling each protip string as a sequence of observed outcomes from a Hidden Markov Model. This modeling decision came about largely due to the fact that HMMs allow for the ordering of words in the original string to be preserved as a result of the Markov

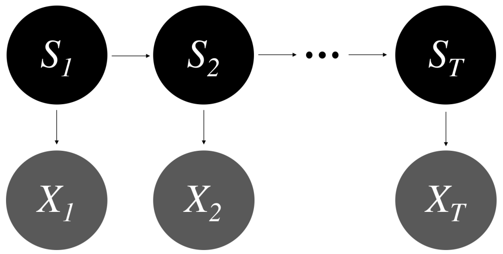

Figure 1: Graphical representation of Hidden Markov Model

property discussed above. This differentiates this modeling approach significantly from a bag-of-words model.

In the context of our problem, each emission X_i of hidden state S_i corresponds to an observed word at the ith index in the protip. Therefore, equation (1) may be more intuitively expressed as:

$$P(\text{protip}) = \sum_{S} P(\text{protip}|\mathbf{S}) \cdot P(\mathbf{S}) \tag{2}$$

The objective is to compute the probability of observing a certain protip string conditioned on which HMM generated it. However, in order to accomplish this, it is necessary to train different HMMs on different corpuses of text, i.e. ethical life protips (LPT) or unethical life protips (ULPT). This way, the probability distributions over the possible output values for hidden states in each respective model are inherently conditioned upon the textual corpus on which the model was trained. Therefore, the semantic interpretation of the probability that a hidden state S_i generated a word X_i is the probability that X_i is observed at that index in the string assuming that the string came from a given corpus of text.

6.1. Classification Algorithm

In employing HMMs to classify protip strings, we generated three HMMs: one trained on LPT, another trained on ULPT, and a third trained on half LPT and ULPT examples, referred to hereafter as the ambiguous HMM. For a given protip string, we fed the string as input to each of the three models and computed the probability of observing that observed sequence of words **X** given the models hidden state sequence **S**. If the highest probability came from either the LPT HMM or the ULPT HMM, the string was classified accordingly. However, if the highest probability came instead from the ambiguous HMM, we took this to be an indicator that the string had an ambiguous ethical valence. Therefore, in this case we employed an alternative ensemble of models to generate a more concrete classification. This model is analogous to the human process of asking friends for a collective opinion when faced by a challenging or unclear situation, and making a decision based on the majority feedback.

It should be noted that the decision to employ this option for classification disambiguation over simply employing a probability threshold is that the latter would have necessitated employing a grid search in order to ascertain the optimal probability thresholds. This option is computationally expensive, requiring the model to be retrained at every possible threshold value. In optimizing for both time and computing resources, we elected to proceed with the use of a third HMM to indicate when it is necessary to employ additional classification methods to assign a label to a string. Additionally, ensemble methods often generalize well to unseen data.

6.1.1 Ensemble Description

In the case that the ambiguous HMM outputted the highest probability of the three HMMs, we defer prediction to an ensemble of three classification models. In selecting the three models to use, we experimented with the following five:

1. Multinomial Naive Bayes
2. Ridge Regression
3. Extra Trees
4. Linear Support Vector Machine
5. Random Forest

The Naive Bayes classifier combines the Naive Bayes probabilistic model with with the maximum a posteriori (MAP) decision rule to output a binary classification.

The Ridge classifier employs a regularized form of regression, that avoids overfitting data by regularizing feature weights to keep them small. Specifically, it minimizes the L2-regularized squared error in constructing a decision boundary.

Extra Trees (short for Extremely Randomized Trees) employs use of randomized decision trees to output a binary classification.

Linear support vector machines work by generating a separating hyperplane that minimizes the L2-regularized hinge loss, which results in a classifier where the distance from the decision boundary to the closest point is maximized.

Finally, Random Forest works by generating numerous decision trees and outputting the label that is the mean prediction of each of the individual trees.

We opted to choose to use an odd number of models in the ensemble in order to avoid ties, given that the final prediction of the ensemble was taken to be the majority prediction. In experimenting with accuracies using 1, 3, or 5 models, subsets with 3 models had the highest overall accuracies. We therefore tested all possible permutations of three of the five models in order to select the highest performing subset.

Model Subset	F1
Naive Bayes, Random Forest, Ridge	0.81832
Naive Bayes, Random Forest, Support Vector Machine	0.81735
Naive Bayes, Random Forest, Extra Trees	0.81613
Naive Bayes, Ridge Regression, Support Vector Machine	0.81583
Naive Bayes, Ridge Regression, Extra Trees	0.82270
Random Forest, Ridge Regression, Support Vector Machine	0.81581
Random Forest, Support Vector Machine, Extra Trees	0.81564
Ridge Regression, Support Vector Machine, Extra Trees	0.81630
Random Forest, Ridge Regression, Extra Trees	0.81661
Naive Bayes, Extra Trees, Support Vector Machine	**0.82221**
Naive Bayes, Extra Trees, Support Vector Machine, Ridge Regression, Random Forest (All models)	0.82026

Table 3: Accuracies from testing various subsets of ensemble models

6.2. Baselines

We built a few baselines to benchmark our model against.

6.2.1 Top K Count

Our baseline model consists of a few phases.

1. Find the 200 most frequent tokens in the ethical and unethical training sets. Call these sets E and U.
2. Let E' be the set of elements in E that are not in U, and U' be the set of elements in U that are not in E. For each training example, we create a feature vector in $\mathbb{R}^2$, where the first element is 1 if the example contains tokens in E' and 0 otherwise, and the second is 1 if the example contains tokens in U' and 0 otherwise.
3. Then, we train a Naive Bayes model on these feature vectors.

6.2.2 Majority

This model just predicts the majority class in the training set.

6.2.3 Length

This model trains on a single feature, the number of tokens in the phrase.

6.3. Oracle

We removed the labels from part of the dev set, and had humans determine whether the provided example was ethical or unethical. This was done on 40 examples.

7. Model Architecture

7.1. Training

The total size of our training set is approximately 32,000 total data points (life protips) with 16,000 coming from r/LifeProTips and 16,000 coming from r/UnethicalLifeProTips. The 16,000 examples from each subreddit were used to train the LPT and ULPT HMMs respectively, and a random subset of 8,000 LPT examples combined with another random subset of 8,000 examples from ULPT were used to train the ambiguous HMM.

7.2. Evaluation

4,000 examples were set aside to use for model evaluation. Performance on this set of data was used to further tweak hyperparameters in effort to increase performance on the test set.

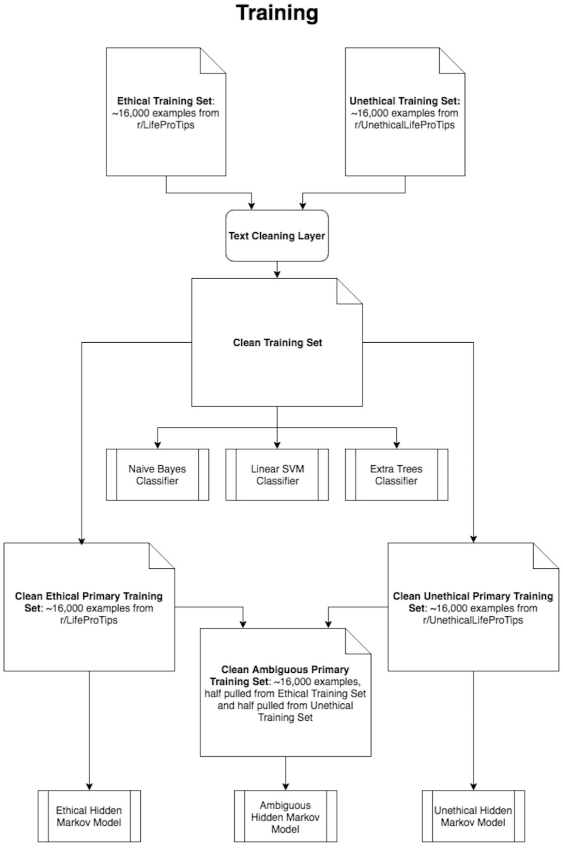
Training
Ethical Training Set: ~16,000 examples from r/LifeProTips
Unethical Training Set: ~16,000 examples from r/UnethicalLifeProTips
Text Cleaning Layer
Clean Training Set
Naive Bayes Classifier
Linear SVM Classifier
Extra Trees Classifier
Clean Ethical Primary Training Set: ~16,000 examples from r/LifeProTips
Clean Unethical Primary Training Set: ~16,000 examples from r/UnethicalLifeProTips
Clean Ambiguous Primary Training Set: ~16,000 examples, half pulled from Ethical Training Set and half pulled from Unethical Training Set
Ethical Hidden Markov Model
Ambiguous Hidden Markov Model
Unethical Hidden Markov Model

Figure 2: Diagram of model architecture (Training)

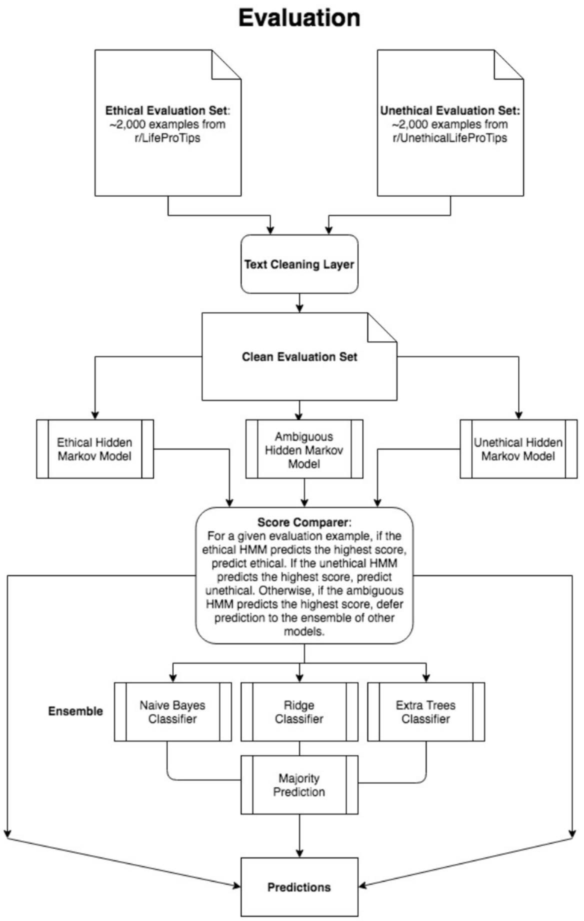
Evaluation
Ethical Evaluation Set: ~2,000 examples from r/LifeProTips
Unethical Evaluation Set: ~2,000 examples from r/UnethicalLifeProTips
Text Cleaning Layer
Clean Evaluation Set
Ethical Hidden Markov Model
Ambiguous Hidden Markov Model
Unethical Hidden Markov Model
Score Comparer: For a given evaluation example, if the ethical HMM predicts the highest score, predict ethical. If the unethical HMM predicts the highest score, predict unethical. Otherwise, if the ambiguous HMM predicts the highest score, defer prediction to the ensemble of other models.
Ensemble
Naive Bayes Classifier
Ridge Classifier
Extra Trees Classifier
Majority Prediction
Predictions

Figure 3: Diagram of model architecture (Evaluation)

8. Hyperparameter Tuning

8.1. Ensemble

Hyperparameter tuning was done via grid search for each ensemble model, to achieve individual optimal performance before ensembling together. For the extra trees and random forest models, it was observed that larger sizes of model led to better performance, while for SVC, NB, and ridge, it was observed that an intermediate parameter selection led to the best performance.

For Naive Bayes, $\alpha = 1, 2$ were tied for the best F1 score at 0.81. This follows logically from the semantic meaning of Laplace smoothing, where smoothing values that are too high cause underfitting, and values that are too low cause erroneously assigned 0 probabilities.

8.2. HMM bias

As mentioned earlier, we use the ambiguous HMM to set a reasonable ambiguity threshold in a more computationally efficient way. That being said, the boundary set by the ambiguous HMM is by no means ideal, so we introduce a parameter, bias, which adjusts the ambiguity threshold to some fraction of the ambiguous HMM's score.

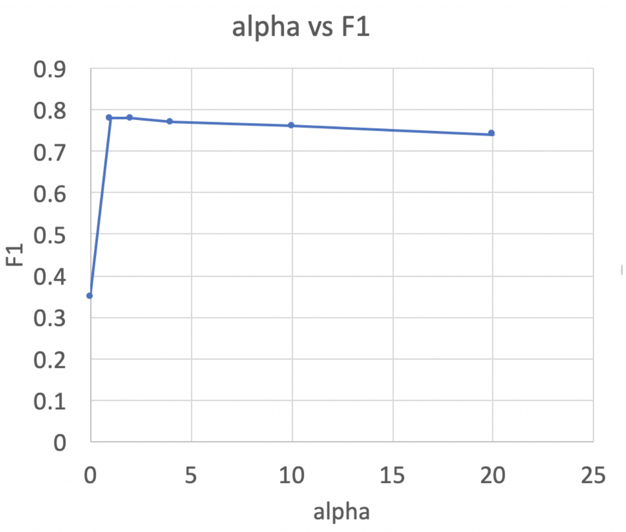

Figure 4: Tuning Naive Bayes alpha parameter

With this parameter, the HMM will defer to the ensemble if $\max(ULPT, LPT) \leq \beta * Ambiguous$, where β is our bias parameter. When $\beta = 1$, we have our original unbiased architecture.

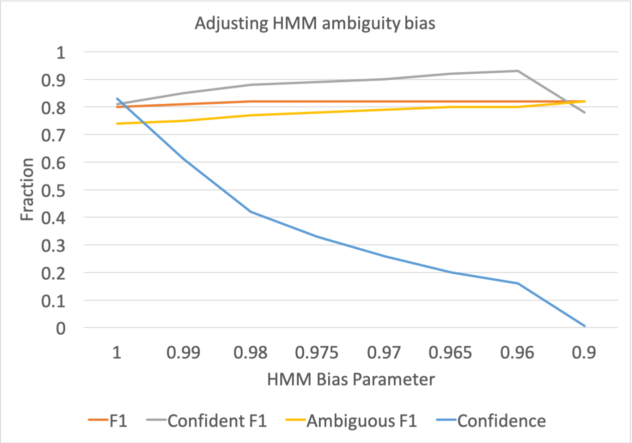

Figure 5: Tuning HMM bias parameter

As we increase the bias toward ambiguity (smaller values of bias parameter), we see that the HMM defers more and more of its decisions to the ensemble (the ”Confidence” line), and the HMM's F1 accuracy on confident values increases, because it only needs to score the examples it is most sure about. Quantitatively, tuning the parameter from 1 to 0.98 increased the HMM's accuracy on confident examples from 0.81 to 0.88. This means that were encouraging the HMM to only make predictions in cases where its quite positive its correct, and to defer to the ensemble otherwise.

The F1 score overall remains constant at 0.82 for all bias ≤ 0.98, so we selected the bias to be 0.98 to maximize the confidence of the HMM at 0.42.

A side note: the reason confident F1 drops off at $\beta = 0.9$ is because the sample size shrunk to a very small number, and those statistics are not representative. With a larger dataset, we would expect confident F1 accuracy to continue to rise with lower bias.

9. Error Analysis

9.1. Samples

By inspecting samples of examples that the model classified incorrectly, we notice patterns of failure that differ based on whether an ethical example was misclassified as unethical or vice versa.

9.1.1 HMM: unethical misclassified as ethical

1. Mislabeled data (is actually ethical)
   - ”cut a notch in your toothbrush , you ll be able to find it in the dark”
   - ”flush the toilet before using it , a slick bowl leaves less skidmarks”
   - ”if your blue pen is leaking ink and spoiling your dress , by creating permanent blue stains , always wear a blue shirt so that it would match the colour of the stains .”
2. Requires domain knowledge
   - ”wrap your pizza in aluminum foil before putting it in the microwave for extra crispy crust .”
   - ”driving with the sun in your eyes is annoying and potentially dangerous . instead , just keep your eyes closed .”
   - ”if you re not happy with your car s sound system , wear headphones while you drive .”
3. Has words that reflect every day tasks
   - ”keep your vacuum cleaner near the door . if you get unforeseen guests , tell them you were in the middle of cleaning up .”
   - ”mark a washer dryer as broken to reserve it for yourself .”
   - ”use a whitening toothpaste when cleaning the rim of the toilet with your roommate’s toothbrush . odor problem solved .”

9.1.2 HMM: ethical misclassified as unethical

1. Related to money
   - ”usually grab something from the gas station before work ? buy in bulk at a grocery store and save tons of money .”
   - ”if you’re a student and want to purchase college apparel , check your local thrift store before paying full price on your campus .”
   - ”if you need some extra money consider selling your plasma .”
2. Has PG-13 words
   - ”get pregnant so you can see who your real friends are”
   - ”don’t kill people because they die when they are killed !”
   - ”it is a long weekend at the end of summer . every cop is looking for drunk drugged drivers .”
3. Mislabeled data (is actually unethical)
   - ”if you receive money on your birthday from someone , either put that in a new card , or use the amount for a gift to the same person .”
   - ”making a new uber eats account will let you get 20 off an order with no minimum amount . free food !”
   - ”if you’re a young pretty girl , crying can get you out of a ticket”

9.1.3 Ensemble: misclassifications

Many ensemble misclassifications were very ambiguous statements.

Examples:

- ”make friends with your bartender . they’ll put more attention into making your drinks timely and excellently .”
- ”if someone insults you , pretend as if you didn t hear it and ask them to repeat it . do it multiple times until they give up or get the point . insults generally lose their effect this way .”
- ”if people keep eating your lunch at work , label it as vegan .”

9.2. Keywords

Comparing the incorrectly classified examples, we can see different patterns in examples incorrectly classified as unethical versus examples incorrectly classified as ethical.

Words associated with examples incorrectly classified as unethical, that appear less in examples incorrectly classified as ethical:

’expensive’, ’posts’, ’large’, ’eat’, ’lost’, **’purchase’**, ’order’, ’likely’, ’break’, ’children’, ’lot’, ’karma’, **’cash’**, ’act’, ”it’s”, **’price’**, ’hotel’, ”they’re”, ’has’, ’service’, ’how’, ’extra’, ’looking’, ’year’, ’end’, ’food’, ’house’, ’local’, **’credit’**, ’offer’, ’restaurant’, ’r’, ’kids’, ’room’, **’card’**, **’buying’**, ’company’, ’first’, ’item’, ’talk’, ’her’, ’gift’, **’pay’**, ’actually’, ’leave’, ’being’, ’simply’, ’now’, ’spend’, ’taking’, ’dog’, ’store’, ’trying’, ’help’, ’;’, ’old’, ’cards’, ’better’, ’friend’, **’paying’**, ’won’, ’sign’

Words associated with examples incorrectly classified as ethical, that appear less in examples incorrectly classified as unethical:

’same’, ’every’, ’using’, ’night’, **’phone’**, **’wash’**, ’clean’, ’come’, ”won’t”, ’0’, ’others’, ’open’, ’change’, ’doing’, **’show’**, ’write’, ’2’, ’without’, **’toothpaste’**, ’bring’, ’play’, **’seat’**, ’front’, ’address’, ’less’, ’10’, ’very’, ’enough’, ’avoid’, **’paper’**, ’through’, **’text’**, ’everything’, ’again’, **’bag’**, ’even’, ’wear’, ’having’, **’email’**, ’turn’, ’else’, ’anything’, ’its’, ’number’, **’door’**, ”you’ll”, ’video’, ’few’, ’fast’, ’making’, **’toilet’**, ’start’, ’may’, ’volume’, ’walk’, ’long’, ’next’, **’water’**, ’gets’, ’used’, ’which’, ’late’

We can see that the model strongly believes words associated with money, like ”expensive”, ”purchase”, ”cash”, and ”credit” are unethical. On the flip side, the model sees every day items as signals for an ethical suggestion: ”phone”, ”toothpaste”, ”paper”, ”door”, ”toilet”.

9.3. Length Distribution

There is potentially bias in the way our HMM handles examples of varying lengths, so we charted the distribution of lengths across correct and incorrect dev set predictions, and found no noticeable difference in the distribution.

The model performs equally well on examples of varying lengths, which is a positive indication that it is actually looking at sentence content rather than following some irrelevant bias.

9.4. Ambiguity

One unique aspect of our model was implementing a notion of ambiguity, which allows the HMM to fall back to an ensemble if it is not confident about its predictions.
We can see clearly here that allowing the HMM to fall back to an ensemble in ethically ambiguous examples (where its F1 was near random at 0.54) allowed it to achieve much higher accuracy at 0.74.

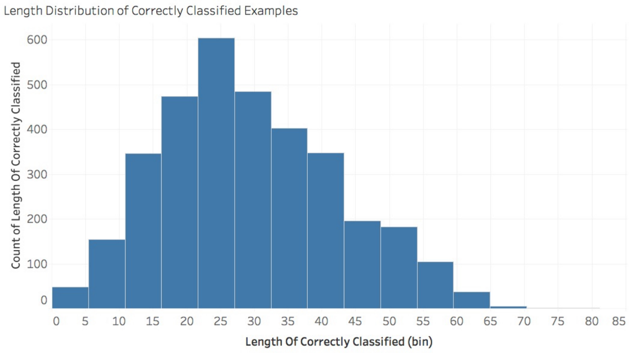

Figure 6: Lengths of correctly predicted examples

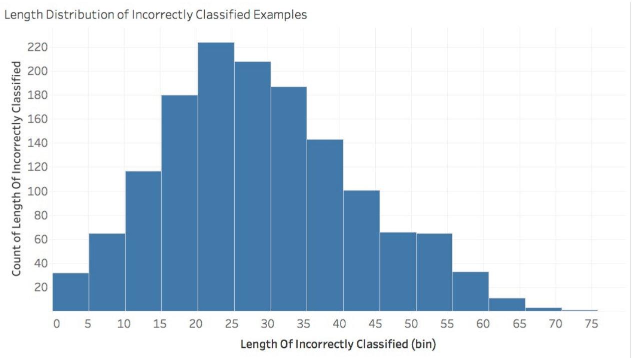

Figure 7: Lengths of incorrectly predicted examples

label	precision	recall	f1	support
0	0.81	0.82	0.81	1727
1	0.81	0.80	0.80	1666
1	0.81	0.81	0.81	3393

Table 4: Performance Analysis for Confident HMM

label	precision	recall	f1	support
0	0.57	0.56	0.56	379
1	0.52	0.53	0.52	339
avg	0.55	0.54	0.54	718

Table 5: Performance Analysis for Ambiguous HMM

label	precision	recall	f1	support
0	0.78	0.72	0.74	379
1	0.71	0.77	0.74	339
avg	0.74	0.74	0.74	718

Table 6: Performance Analysis for Ensemble Fallback

This result encouraged us to decrease the ambiguity threshold required to fallback to the ensemble, since the performance of the ensemble is clearly much better in ambiguous situations, and it is reasonable to assume that we can achieve performance gains through better generalization.

Examining some of the errors made by the HMM that were near the ambiguity threshold, we see that the model made poor calls on examples that were semantically near the margin.

For example, the HMM incorrectly classified the following two statements as unethical, when they are labeled as ethical by Reddit:

- ”if an ad asks you to select your age group , select 13 and under . most of the time you won t get an ad since there aren t many ads created for that age group .”
- ”if you’re buying a car from a dealership and know exactly what you want , tell the dealer you’re torn between that and another car , but leaning towards the other car . the dealer , wanting to close the deal , will reveal the negative aspects of the car you actually want to get .”

By deferring these unclear examples to the ensemble, the overall model can achieve better performance, and the HMM can stick to outputting classifications on examples that it is confident on.

10. Results

Model Type	F1
HMM with ensemble backoff, no bias	0.78
HMM with ensemble backoff, bias=0.98	**0.82**
Ensemble only	**0.82**
HMM only	0.74
Ensemble + HMM equal voting	0.81
Human	0.805

Table 7: Final results

Model Type	F1
Top K	0.61
Majority	0.33
Length	0.51

Table 8: Baselines

The best two models we experimented with were HMM with ensemble backoff (bias parameter 0.98) and ensemble only. All tested models performed significantly above baselines.

Adding ambiguity and adjusting bias both played a large role in improving performance of the HMM backoff model, each improving F1 by 4 percentage points.

The HMM with backoff was able to perform with a higher score than human oracles, and thus it is likely that creating models with higher accuracy will be difficult until a more universal definition of ethics is codified.

11. Conclusions

Determining ethics in a statement is, in many ways, a profoundly human endeavor. In applying computational models to the problem, there are two important points to keep in mind:

1. The ways in which computational models consider ethics is fundamentally different than processes of human cognition.
2. Computational models for ethics should be interpretable as well as high performant, because ethics is a value-based problem, and making sense of the values being learned by the machine is, to an extent, just as important as correctly classifying statements.

To this end, we have created an ambiguity-based hidden Markov model, which performs as well as a large optimized ensemble, while affording researchers the ability to examine hidden states to understand what the model is internalizing, embed new knowledge into the model in the form of priors, and capture statistics about when the model is unsure of its decision and is forced to defer to a more traditional architecture.

By creating interpretable computational ethics models, humans can not only extend ethical decision making to a larger scale, but also, through the examination and pertubation of the internals of such models, potentially come to a new and more data-driven viewpoint on ethics itself.

References

[1] W. Wallach, “Robot minds and human ethics: the need for a comprehensive model of moral decision making,” *Ethics and Information Technology*, vol. 12, pp. 243–250, Sep 2010.

[2] L. Liu, D. Luo, M. Liu, J. Zhong, Y. Wei, and L. Sun, “A self-adaptive hidden markov model for emotion classification in chinese microblogs,” *Mathematical Problems in Engineering*, vol. 2015, p. 18, 2015.

[3] M. Kang, J. Ahn, and K. Lee, “Opinion mining using ensemble text hidden markov models for text classification,” *Expert Systems with Applications*, vol. 94, p. 218227, 2018.

[4] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” pp. 1532–1543, 2014.