A Machine Learning Approach to Solving Ethical Relativism

Taylor Olson[1]

University of Northern Iowa, Cedar Falls, IA
United States

Abstract. The lack of universality in ethics presents problems in areas well beyond that of itself. I explore an approach involving data science and machine learning, to reduce every culture's moral codes into an atomic list of laws. With the emergence of social media, we have direct access to user opinion. By analyzing natural language in tweets, statistics can be gathered on user positions on popular ethical issues. This paper presents a multi-class machine learning classifier that can be used to find the distribution of users for, against, or neutral to an issue that involves moral reasoning. With this distribution, the system can then build a reduced code of ethics that the majority of humanity adheres to.

1 Motivation

Ethical relativism is the theory that moral/ethical positions are not objective, but are entirely dependent upon one's culture and past experiences. For example, most western cultures view genital as fundamentally wrong, while genital mutilation is prevalent in the African countries Yemen, Iraqi Kurdistan and Indonesia [5]. While some philosophers reject the idea of ethical relativism, the existence of multiple code of ethics cannot be ignored as it brings about challenges in many areas. Bostrom illustrates one major challenge in the area of artificial intelligence with his Value Choosing and Loading problems [6]. With the emergence of intelligent systems, we must determine how the systems will gain their values and what values we wish them to gain. But what culture's values do we choose? A wrong decision could be catastrophic. Among the many other challenges that emerge from ethical relativism is determining when to interfere with foreign affairs. A clearly relevant problem today with no clear solution. This project presents a proposed solution to identifying a universal moral code that all of humanity adheres to, by using machine learning to analyze the positions of users on social media on popular ethical issues.

2 Methodology

The proposed solution involves a five step process. First, gather a large dataset of tweets on popular ethical issues. Second, build a machine learning classifier capable of deeming a tweet as **For**, **Against**, or **Neutral** to each issue. Third, continually gather more tweets and use the classification model to classify each tweet into one of the three classes. Fourth, when a sufficient amount of tweets have been collected and classified, gather and analyze statistics of people's positions on each issue (percentage of people for, against, and neutral to each issue). Lastly, use the statistics to build a reduced code of ethics. Assuredly, there are limitations and challenges involved with this process. Therefore, I have provided a list of recognized arguments, responses to those arguments, as well as proposed solutions in the Conclusion and Future Work section.

2.1 Reducing Ethical Issues

To fully analyze each ethical issue, stances on an issue need to be reduced to an atomic list of laws a priori. For example, 'against abortion' could map to 'thou shall not kill'. In hopes to not beg the question, this may need to be further reduced to 'thou shall not terminate the unborn'. Though this leaves room for further disagreement about the state of being of the 'unborn'. In hopes of not entering into an endless loop of semantic argument, the following reduction map in figure 1 will serve as satisfactory.

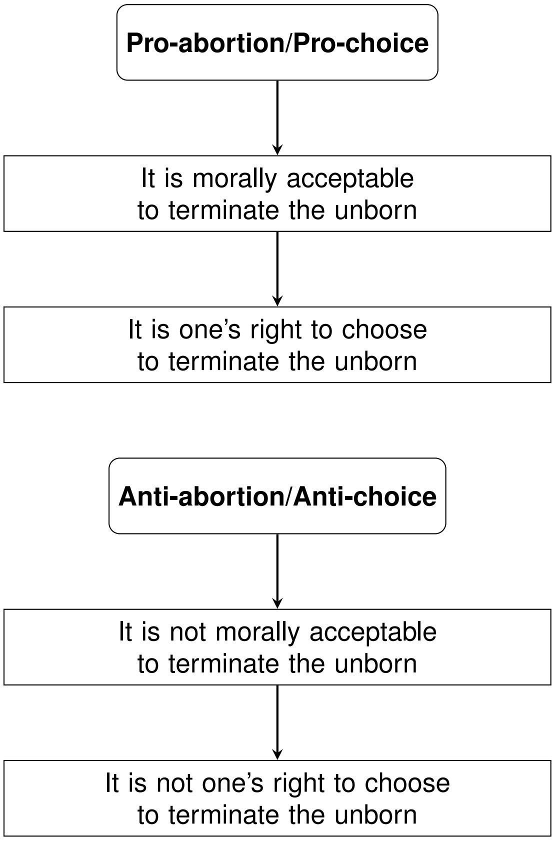

Pro-abortion/Pro-choice
It is morally acceptable to terminate the unborn
It is one’s right to choose to terminate the unborn
Anti-abortion/Anti-choice
It is not morally acceptable to terminate the unborn
It is not one’s right to choose to terminate the unborn

Fig. 1. Reduction mapping of stances on abortion

2.2 Data Gathering

A dataset of 10,000 public tweets was built by using a simple keyword search for ’abortion’. Three labels were recorded: users Twitter handle, tweet text, and geolocation. For sake of time, only 3,141 of the 10,000 tweets were manually labeled into three different classifications: **For, Against, Neutral**. The labeling of ”For” and ”Against” was determined with high scrutiny. Meaning, only tweets that were clearly Pro-Choice/Pro-Abortion were labeled as ”For”, otherwise ”Neutral”. Similarly for ”Against”. Table 1 shows the resulting distribution of classes in the manually labeled data set.

Table 1. Distribution of classes in dataset

Classes	Tweets
For	562
Against	801
Neutral	1,778

2.3 Preprocessing

Preprocessing of the dataset consisted of 6 steps:

1. Removing twitter handles
   Twitter handles were removed from all tweets not only for privacy purposes, but also due to the lack of information they provide. The sentament of a tweet rarely (if ever) changes due to who the user ”@’s”.

2. Removing non-English
   Non-English was removed primarily for simplicity purposes. A translation system (from language X to English) could have been used, but this adds an unnecessary level of complexity. Therefore, the scope of this project involved analyzing only English tweets from the United States.

3. Removing numbers
   Clearly numbers give text little meaning (at least for the scope of this project). Therefore, all numbers are simply removed.

4. Lowercasing
   To normalize all text, every character is mapped to it’s lowercase counterpart. This ensures that all words, regardless of case, are mapped to the same string. For example, $\{$’HATE’, ’Hate’$\} \rightarrow$ ’hate’. This is absolutely necessary because a Bag-of-Words model is used for feature extraction.

5. Removing punctuation
   Tweets rarely contain punctuation. Thus, in hopes to normalize all tweets, all punctuation was removed (excluding hyphens).

6. Removing all occurrences of the term ’abortion’
   Because tweets were gathered by using a keyword search for the string ’abortion’, all

tweets will contain it. So this string (and every form of it) serves no purpose in separating the dataset into classes. Therefore, all strings containing the sub-string 'abortion' were removed (#abortion, abortions, etc.).

7. Feature Extraction via Bag-of-Words
   One of the most common methods of feature extraction was used, the Bag-of-Words (BOW) approach. With this method, "we look at the histogram of the words within the text, i.e. considering each word count as a feature" [1]. Two variations of features were used: binary/non-binary, and n-gram (gram = word, n = 1, 2, 3). As a result, each tweet was represented as an N-dimensional vector (N = number of words in vocabulary or number of n-grams) with the value at each n-th dimension being a word/gram count or binary value.

2.4 Training Classifiers

For training, the manually labeled dataset was split into a training set of 2,000 tweets and the entire dataset of 3,141 tweets was used for testing. The distribution of classes in the training set is shown in table 2 below.

Table 2. Distribution of classes in training data

Classes	Tweets
For	600
Against	800
Neutral	600

Three different supervised classification techniques were then implemented and tested:

1. Multinomial and Bernoulli Naive Bayes
   Naive Bayes classifiers take a probabilistic approach to classification, where the probability of a tweet **d** being in class **c** is given by:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \text{ [2]}$$

   where $P(t_k|c)$ is the conditional probability of seeing term t_k in class c. When classifying tweets, the Bernoulli model uses binary occurrences (0 if a term occurs in a tweet, 1 otherwise), while the multinomial model records multiple occurrences. [2].

2. Support Vector Machine with rbf and linear kernels
   "In support vector machines the decision boundary is chosen to be the one for which the margin is maximized" [3] Where margin is defined as the perpendicular distance to the closest point x_n from dataset. The goal of SVM is then to optimize **w** and **b** to maximize the margin, which is given by:

$$\arg\max_{w,b} \{ \frac{1}{||w||} \min_n [t_n(w^T \phi\ (x_n) + b)]\} \text{ [3]}.$$

3. K-nearest neighbors
   This classification model uses a similarity measure to get similarity scores between the tweet in question and all tweet vectors in the training dataset. It then uses a ranking system and gets the highest K amount of similarity scores. Of the highest K tweet vectors, their corresponding classes are analyzed and the class with the most "votes" is assigned to the tweet. A K-value of 1, 2, and 3 was tested and Euclidean distance (equation shown below) was used for the similarity measure.

$$d(p,q) = \sqrt{(p_1 - q_2)^2 + ... + (p_n + q_n)^2}.$$

2.5 Classifier Performance

For testing the Naive Bayes classifiers, K-fold cross validation was used. In this method, The dataset is first split into k parts, and the holdout method is repeated k times. "Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed." [4]

For testing the Support Vector Machine and K-nearest neighbor classifiers, the trained models were tested only once on the entire manually labeled dataset and the average error was computed as: amount correctly classified / total amount.

As stated in the preprocessing section, two variations of feature extraction were used and compared: binary/non-binary, and n-gram (gram = word, n = 1, 2, 3). The resulting accuracies for all classifiers are shown in figure 2.

K-nearest neighbors, with a K value of 1, performed the best with an accuracy of 95.25% during testing. However, when tested on other sample datasets of tweets, it performed incredibly poorly. This was most likely due to the fact that the K-value of 1 over-fits the model to the gathered data. Therefore, the next best performing model with a testing accuracy of 86.60% was used, a SVM with a linear kernel. The SVM model performs well when given tweets that are clear. However, when given tweets containing sarcasm or logical reasoning, it often mis-classifies. Potential solutions are discussed in the Recognized Problems and Future Solutions section. Three trivial and one non-trivial example of tweets being classified by the SVM model are shown in figure 3 on the second to last page.

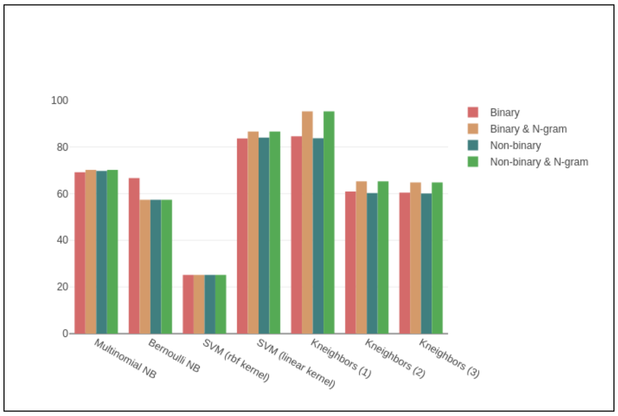


Fig. 2. Classification model results

2.5.1 Interface Prototype

To help constantly monitor statistics and easily analyze the classifier's performance, an interface was built on top of the system. Images of the interface are provided on the last few pages. Figure 4 shows a screen shot of the table view that allows users to view tweet text and classifications, while figure 5 shows the map view that allows a user to analyze clusters for classifications of tweets. In both views, users can enter in an amount of most recent tweets they wish to view and press load tweets to view the percentage breakdown of the three classes.

3 Conclusion and Future Work

Analyzing over 2,000 tweets, a common trend has been found in U.S. positions on abortion. Around 40% of the U.S. tweets are consistently anti-abortion/anti-choice and around 20% are pro-abortion/pro-choice, the rest being neutral. With the reduction map as defined, we can then conclude that the majority of the U.S. believes that it is not one's right to choose to terminate the unborn. I have shown that data science and machine learning, if extended outside of the U.S., can be used as a culturally independent approach of analyzing ethics and is therefore a potential solution to ethical relativism. In the future, a larger list of ethical issues should be considered. Issues such as gay marriage, euthanasia, and the death penalty will be analyzed using the same proposed methodology, which will result in a more comprehensive atomic code of ethics.

3.1 Recognized Problems and Future Solutions

Problem: Positions on each ethical issue would be best analyzed by examining user reasoning and motivation for taking a particular stance, rather than the stance itself. For example, consider the two common and opposing beliefs: a fetus is a living being, and a fetus is merely a cluster of cells. One's stance on abortion that believes the former can not be reduced in the same way of that of the latter. However, this level of reasoning and motivation cannot be captured by looking solely at word frequencies within tweets.

Solution: To solve the issue of reasoning rather than position, an advance argument analysis system could be built to extract user arguments and steps of reasoning. A system similar to the argument extractor presented in the paper *Argument Mining: Extracting Arguments from Online Dialogue* [7]. Said system would be able to deconstruct an argument within a tweet into it's predicates and conclusions. This would allow the classification algorithm to analyze user assumptions and reasoning rather than just a binary stance on the issue.

Problem: Analyzing only the U.S. still suffers from ethical relativism, as it only accounts for a subset of all cultures.

Solution: In hopes to fully capture a vast amount of different cultural opinions on popular ethical issues, data will surely need to be gathered and analyzed from outside of the United States.

Problem: Assuming that analyzing tweets alone encapsulates all of the population's beliefs on each issue.

Solution: With enough storage and computing power, a system could theoretically be built that analyzes information from multiple online sources. The system would predict an individual's position on ethical issues by scanning personal websites, Facebook, and even research papers. This solution is still limited to societies with strong online presence, but it covers much more of the population than Twitter alone.

Resources

1. Python 3.6.1
   Usage: Main programming environment
2. tweepy 3.5.0
   Author: Joshua Roesslein
   Usage: Gathering tweets
   Home-page: http://github.com/tweepy/tweepy
3. pandas 0.20.1
   Author: The PyData Development Team
   Usage: Data storage and manipulation
   Home-page: http://pandas.pydata.org
4. gensim 2.3.0
   Author: Radim Rehurek
   Usage: Feature Extraction
   Home-page: http://radimrehurek.com/gensim
5. scikit-learn 0.19.1
   Author: Andreas Mueller
   Usage: Classification models
   Home-page: http://scikit-learn.org
6. Plotly dash 0.18.3
   Author: chris p
   Usage: Interface and plotting
   Home-page: https://plot.ly/dash

References

1. Yoav Goldberg, *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, 2017

2. Christopher D. Manning, Prabhakar Raghavan, & Hinrich SchÃ$\frac{1}{4}$tze, *Introduction to Information Retrieval*. Cambridge University Press, 2008

3. Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006

4. Jeff Schneider and Andrew W. Moore, *Pattern Recognition and Machine Learning*. Carnegie Mellon University, February 1, 1997

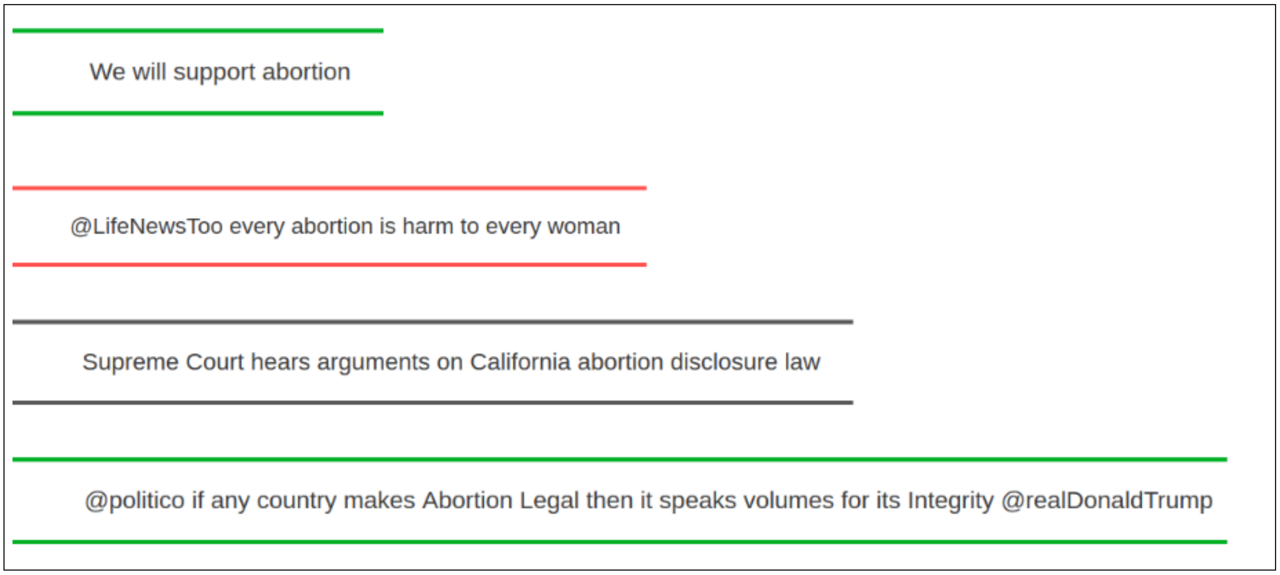
5. Emma Batha *What is genital mutilation? Where does it happen?*. Reuters, Thomson Reuters, 30 January 2017

6. Nick Bostrom *Superintelligence Paths, Dangers, Strategies*. Oxford University Press, 2014

7. Reid Swanson, Brian Ecker, & Marilyn Walker *Argument Mining: Extracting Arguments from Online Dialogue*. UC Santa Cruz

Acknowledgement

I would like to thank Dr. Mark D. Ecker for his guidance and feedback throughout the duration of this project.



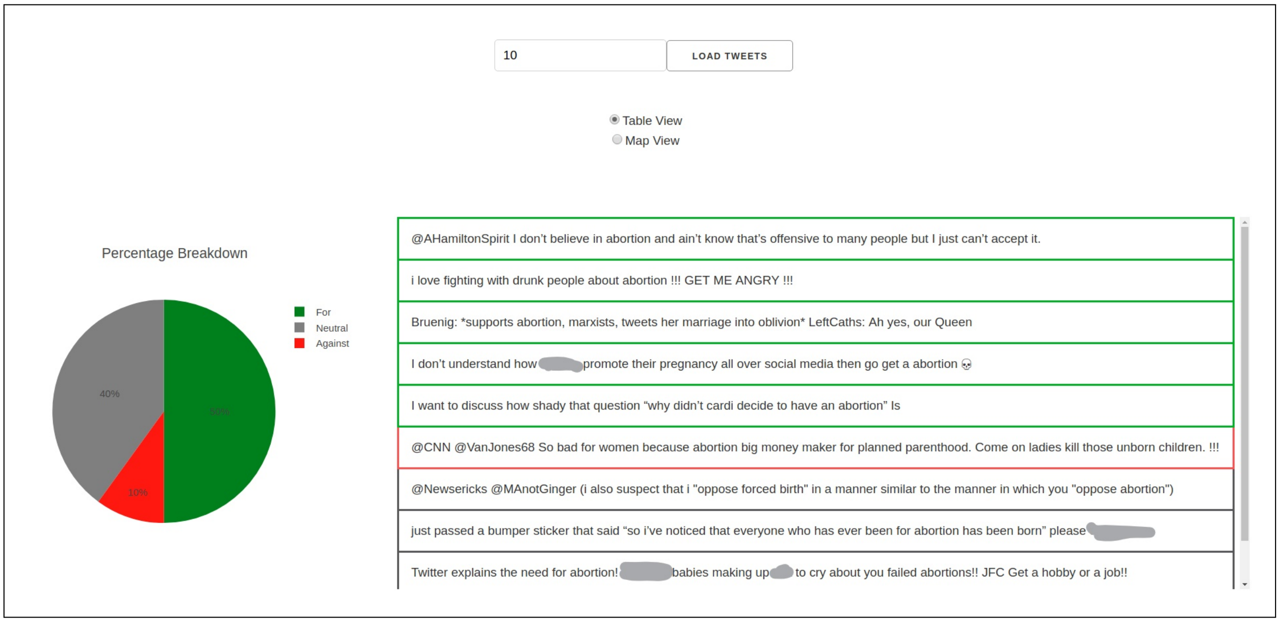
We will support abortion

@LifeNewsToo every abortion is harm to every woman

Supreme Court hears arguments on California abortion disclosure law

@politico if any country makes Abortion Legal then it speaks volumes for its Integrity @realDonaldTrump

Fig. 3. SVM classification examples. Green: Pro-abortion, Red: Anti-abortion, Gray: Neutral



@AHamiltonSpirit I don’t believe in abortion and ain’t know that’s offensive to many people but I just can’t accept it.

i love fighting with drunk people about abortion !!! GET ME ANGRY !!!

Bruenig: *supports abortion, marxists, tweets her marriage into oblivion* LeftCaths: Ah yes, our Queen

I don’t understand how [redacted] promote their pregnancy all over social media then go get a abortion 💀

I want to discuss how shady that question “why didn’t cardi decide to have an abortion” Is

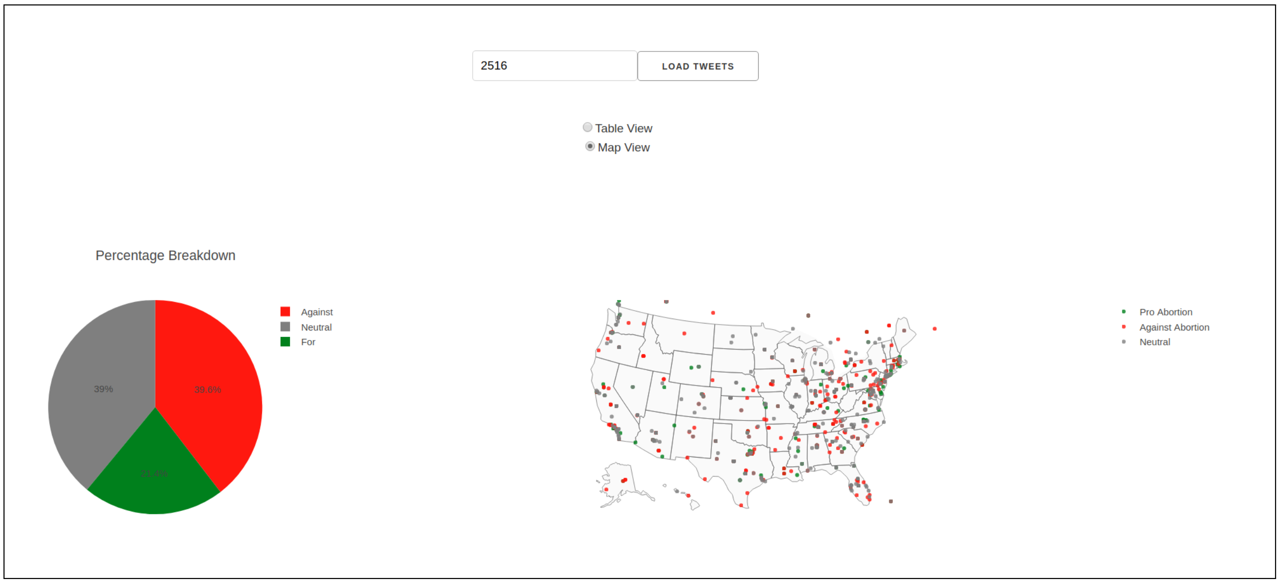
@CNN @VanJones68 So bad for women because abortion big money maker for planned parenthood. Come on ladies kill those unborn children. !!!

@Newsericks @MAnotGinger (i also suspect that i "oppose forced birth" in a manner similar to the manner in which you "oppose abortion")

just passed a bumper sticker that said “so i’ve noticed that everyone who has ever been for abortion has been born” please [redacted]

Twitter explains the need for abortion! [redacted] babies making up [redacted] to cry about you failed abortions!! JFC Get a hobby or a job!!

Fig. 4. Interface, table view



2516 LOAD TWEETS

○ Table View
● Map View

Fig. 5. Interface, map view