(A)moral Machine: exploring the creation and application of an automated moral classifier

Livia Teernstra

August 26, 2015

Media Technology Graduation Project
Leiden University
Supervisors: Dr. Fons Verbeek, Dr. Peter van der Putten, Dr. Liesbeth Noordegraaf-Eelens
Email: livia@helloliefje.com

Abstract

This study tests the possibility of teaching a computer program to detect elements of morality. It uses a novel application of supervised machine learning to identify the presence of underlying moral perspectives in the social media application "Twitter". Moral Foundations Theory sets the framework for Multinomial Naïve Bayes and Maximum Entropy models to classify "Tweets". The Naïve Bayes model achieved a similar level of classification accuracy as human coders, outperforming Maximum Entropy. The successful classifier was then applied to unlabelled Tweets regarding the Greek exit from the Eurozone. The results show that Tweets can be classified into moral foundations to examine changes in people's moral perspectives over time.

1 Introduction

In the past decade, there has been increasing momentum for an interdisciplinary approach to social sciences research. Many social scientists outline a need for academic collaboration, especially with computer science approaches, because the latter displays an ability to access and analyse large amounts of data with a relatively small amount of material resources [1]. Today, the marriage of these disciplines usually draws its data from the Internet. Internet media provides a public sphere where people can engage and converse without the traditional limitations of space and time [2]. The instantaneous nature of internet communication provides an immediate outlet for emotions, opinions, information and interactions, tinted with moral perspectives.

"Twitter", a microblogging social media platform founded in 2006, embodies these forms of behaviours. This platform is favoured as a source of data by

researchers for two reasons. Firstly, communication is limited to 140 characters and can provide basic insights on public opinion [1, 3]. Secondly, its public application processing interface (API) allows for ease of data retrieval. Given that ethics are noted to be a driving force of many human interactions, one can expect expressions of moral values to be found in Twitter communication (also known as "Tweets").

This would potentially make Twitter a promising experimentation ground for interdisciplinary research on morality. However, most interdisciplinary research examines social interaction through metadata, such as ratings, "shares" and hyperlinks [4–6]. Previous research focuses on the diffusion of information rather than the content. Even when content has been analysed, this has mostly been focused on commercial products or political sentiments [7, 8]. This paper submits that based on the nature of Twitter interaction mentioned above, moral perspectives could usefully be analysed using similar opinion mining techniques.

This study aims to investigate which moral perspectives are expressed in Tweets using a social psychology approach called Moral Foundations Theory (MFT), as exemplified in the recent writings of Haidt, among others [9, 10]. MFT provides clear moral categories (called foundations) which can be applied to empirical testing. This is the first study to use a novel application of machine learning to detect these foundations. Hence, the research question raised is: To what extent can supervised machine learning detect moral foundations in Twitter communication?

Austerity is a good discussion in which to test moral values because it has often been criticized as a problem of moral hazard [11]. One particularly controversial political issue that has been in the forefront of European news since 2012 regards austerity measures in the Eurozone. More specifically, this study will explore public opinion on the Greek exit of the Euro (the 'Grexit'). Many analysts note that the "irresponsible behaviour" of (southern) governments is the root cause of the European financial crisis [12]. Consequently, this discussion is framed in a moral light, where 'good' and 'bad' nations are distinguished. This austerity dispute will be used to contextualize the methodology since it has the potential to engage all moral foundations, depending on the individuals frame of reference.

However, it should be noted that the moral undercurrents of the Grexit discussion are not primary concern of this study. Instead, the social and scientific relevance for this research lies in the novel application of MFT to categorizing Tweets. Firstly, most attempts to automatically classify Tweets use dictionary based pre-processing techniques, meaning that large word lists grounded in psychological theory must be first created and verified [7, 13, 14], as opposed to being discovered automatically by machine learning. Secondly, previous research applying MFT to rhetoric analysis focuses on the frequency of words in these created dictionaries to create moral loadings for texts [8, 15, 16], but the relative importance of those frequencies for detecting certain moral foundations is not derived from evidence. Lastly, the texts analysed are authored by opinion leaders, such as news media or "bloggers", rather than the general public.

Primarily, this is the first study which attempts to classify Tweets into Moral

Foundations. It will use tested machine learning techniques to explore moral expressions in a natural setting, through the use of the Twitter platform. Predominantly, the study will determine if supervised models are able to classify morality into MFT categories at an acceptable accuracy. Using machine learning algorithms has academic relevance for MFT, since they can automatically determine lexical indicators for each foundation, without the need to create a dictionary beforehand. This method also has further relevance for MFT, since it will also include an additional moral category, which will be discussed further in section 2. Thus, lexical indicators from this additional foundation can be used as a starting point for further research in value expression of moral foundations.

Specifically, this study aims to answer the research question through using Naïve Bayes (NB) and Maximum Entropy (ME) classifiers in supervised machine learning, which will be further explained in Section 3. As these are supervised learning models, training and testing sets will be used to determine their accuracy. Following extensive testing, the model will be applied to examine the distribution of MFT values to unlabelled Tweets. In a practical sense, the tested algorithm will be able to classify Tweets by their dominant moral underpinning.

To better understand the output, one needs to first understand the motivations behind the classification of the data into moral categories. As such, MFT will be briefly outlined in section 2.1. This will be followed by how these foundations are present within the context of the Grexit. Then, section 3 discusses the methodology of data collection, pre-processing and analysis. Following this, experiments evaluating the classifiers and their results will be presented in section 4. Then, the best performing classifier will be applied to unlabelled Tweets, resulting in a time-series comparison of moral discussion on the Grexit. Finally, results and limitations will be discussed in section 5, followed by the final conclusions of the study.

2 Literature Review

In section 1, it was outlined that interdisciplinary research methods can be used to examine moral reasoning. But where does this moral reasoning come from? Contemporary moral research asserts that the root of our morality lies in a combination of biological and environmental factors, which motivate our decision making [8]. These deep-seated motivations can serve different social functions, and the presence of moral undertones in rhetoric can impact moral and political world views [17]. This section first provides an overview of MFT, showing how these foundations can be applied to the case study of the Grexit. Then, previous research that has applied MFT to supervised machine learning will be noted, followed by other machine learning uses in analysing Twitter data.

2.1 Moral Foundations Theory

Intrinsic, cognitive responses lie at the base of MFT, where the authors use these responses to explain the variation in human moral reasoning across cultures [18]. Otherwise said, the theory states that there is an innate and universal morality which transcends cultural boundaries. This universal morality can be categorized into different foundations, which can be thought of as 'moral building blocks'. Each foundation is fostered within cultures, constructing narratives, virtues and institutions.

How these foundations are built upon differs between groups, where some may emphasize one foundation over another. Yet it does not also discount that moral values can be simultaneously held by individuals and societies. These values can not be objectively ordered in terms of importance, and may be conflicting with one another. Due to this, MFT asserts a pluralistic stance, postulating that all humans are driven by the same innate moral intuitions. However, the degree of importance and emphasis on each value differs between individuals and societies [10]. Clearly, MFT provides a strong starting point for examining moral value judgements on controversial, international topics. Briefly described, the five foundations are:

1. Care / Harm - The desire to cherish protect others, identification of a victim and sympathy with him.

2. Fairness / Cheating - The notions of justice and rights, applied to shared rules in a community. Relates to reciprocal altruism.

3. Loyalty / Betrayal - Relating to 'in-groups'; friends, family, community, as well as showing virtues of patriotism.

4. Authority / Subversion - Submission to and respect for legitimate authority and traditions.

5. Sanctity / Degradation - Stems from feelings of disgust and and contamination. Relating to the virtue that 'the body is a temple', and should not be defiled.

Although these are five foundations which form the basis of MFT, there has been some discussion and strong suggestions to include Liberty and Oppression as a sixth value. It can be described as such:

6. Liberty / Oppression - The resentment of tyranny and desire for autonomy. This is often in tension with the foundation of 'Authority'.

In the context of this research, these six basic foundations of MFT will be used to classify Twitter data. Liberty has previously been included in the model for other politically driven studies, and some researchers have expressed their desire to endorse liberty as a moral value [10]. Thus, the foundations outlined are not set in stone and can be flexible, depending on the research space in which it is contextualized.

As mentioned in section 1, research of MFT value expression has lately focused its applications in political ideology, making it a useful framework to apply to the Grexit. Talks of a Greek withdrawal from the Eurozone monetary union arose in 2012, calling for Greece to deal with its increasingly unmanageable public debt. The main argument is that leaving the Euro (and consequently reintroducing the Drachma) will boost the Greek economy through improving exports and tourism, as well as discourage costly imports [19].

In January 2015, the Greek economist Yanis Varoufakis was appointed as the Greek minister of Finance, but quickly left the position in July 2015. Varoufakis' actions in his position as the minister of finance have led to the Grexit taking centre stage in the Eurozone crisis discussion for the first half of 2015 [20]. Due to his economic background, Varoufakis conforms to the standard belief that those in the economic and financial sectors make decisions solely based on empirical evidence, separating personal moral judgements from their work [21]. Notably, Varoufakis openly rejects a moral narrative in political discourse [22].

Yet there are many moral currents that meander through the debate. For example, it is argued that a Grexit could further hurt the image of the Eurozone, and eventuate in the alignment of Greece with other Non-EU states [20]. This is woven into the foundation of 'loyalty', where Greek loyalty may shift away from the European union. 'Care' is also an important moral driving force in this discussion. One of the key opposing arguments for the Grexit is that the Greeks will be impoverished, leading to nationwide dissatisfaction, which can then result in civil unrest [23]. Interestingly, the focus is on the outcome of an impoverished society, rather than the suffering of the people within it. Due to the clear moral underpinnings in arguments for a Grexit, the situation provides a good experimentation ground for the observation of moral value expressions by the general public.

The application of MFT to the Grexit, as well as example Tweets for each foundation can be found in Table 1. The examples in Table 1 show that each foundation is clearly present in Tweets about Greece and the Euro. Most foundations can be seen multiple angles, with different virtues for Greek and non-Greek perspectives. For instance, the moral importance of liberty is expressed through desire for Greece to leave the Euro. Whether this is a virtue or a vice is unclear, since it could be a virtue for Greece (freedom from the tyranny of Eurozone austerity measures) or for the other European nations, where they would be free from providing further loans to Greece. In addition, 'care' could be related to the struggle of the Greek economy, or the Eurozone as a whole, and the victim could be either party.

Clearly, the moral foundation which drives ones opinions can have many roots. They can stem from society at large, smaller communities, or ones innate moral intuitions. As such, this study asserts no preference for a specific moral standpoint, as its main focus is the use of supervised machine learning in moral classification. All in all, MFT provides a framework for identifying moral expressions, which can be applied to Twitter discussion regarding the Grexit.

	Related Concepts	Example Tweet
Authority	Authority figures (ie: IMF)	• "Greece says Euro zone approves reform plan" • "German elites are willing to let the Euro crash to guarantee their own political survival"
Care	Implying a victim in the Eurozone situation, which groups should be protected	• "European control of the IMF is helping Greece" • "Greece runs out of funding options despite Euro zone reprieve"
Fairness	(Un)just treatment of nations	• "Greece forced to sell assets and cut spending to pay back debts to EU. • "It's easy for the Dutch to go hard on Greece"
Liberty	Liberty for Greece or the other Eurozone nations	"Greece needs a path out of the Euro" • "Greece really might leave the Euro"
Loyalty	Patriotism for a certain nation or group of people	• "If I had to choose between #Greece and #Germany, I know which way I'd go..." • "Greece may stay in the Eurozone for the time being there are no guarantees it can become a responsible member."
Sanctity	The status-quo being the 'embodied temple' which can be defiled by economic actions	• "There really is no space inside the Euro for a radical left government" • "the four-month extension on the Greek debt lowers the risk of Greece leaving the Euro zone"

Table 1: Moral Foundations in the Grexit

2.2 Machine learning approaches to text analysis

There are several machine learning approaches to text analysis, using supervised and unsupervised methods. Supervised learning is appropriate to use for creating a program that can automatically classify data, such as the classification of Tweets by their dominant moral foundations. But there is little research using supervised machine learning to classify data based on MFT. Since there is no current research using MFT in Twitter analysis, this subsection outlines previous machine learning research based on either MFT, or Twitter.

As mentioned in section 1, dictionary based approaches are often used in Tweet classification. In line with this trend, a Moral Foundations Dictionary (MFD) was developed by some of the founders behind MFT. This dictionary gives linguistic indications for the five basic moral foundations (hence, 'liberty' is excluded). The dictionary is separated by virtue and vice terms for each foundation. It is usually used as an add-on to the Linguistic Inquiry Word (LIWC) count program [17]. LIWC is one of the most widely used tools for text analysis, especially sentiment analysis [14,16]. It is also commonly used for Tweet classification. Yet, there is no current research which combines the MFD and LIWC to automatically learn to detect moral foundations in Tweets.

Instead, current textual research using LIWC and the MFD focuses on specific, predefined keywords (and collocations of those words) to examine the extent of the presence of moral foundations. It has been applied in analysis of long texts such as news articles and web blogs, where rhetorical moral assessments were assigned to each text [8, 16]. For example, research on the Ground Zero Mosque showed that blog authors showed more lexical similarity amongst virtuous terms for the foundations care, fairness and authority [16]. One can then gather that expression of the other foundations may be constructed differently amongst cultural groups. Due to the differences in textual expressions of moral opinion, dictionary based approaches can be problematic when drawing conclusions about moral reasoning. All in all, the use of the MFD in text analysis is in its infancy, and there is notable room for a variety of applications and refinement of the indicators.

Presence of ones moral reasoning is not only limited to longer texts. As seen in section 2.1, citizens use Twitter to actively engage in civic matters with moral undertones. Despite this easy access to rich communication data, politically related Tweets are often examined in network analysis to study the spread of misinformation during elections [5,14,24–26]. When communication is examined, it is usually in an opinion mining context. Most commonly, opinion mining is used to predict the result of elections [14, 24]. Consequently, there is little research that has examined Tweets further than looking at positive or negative classification of sentiments. Since previous research has focused either on sentiment or information networks, this will be the first study to attempt to classify moral foundations in Tweets using this method.

3 Methodology

This section covers the data gathering and processing techniques employed, followed by the difference between Multinomial Naïve Bayes and Maximum Entropy classifiers. It will also explain the method for evaluating the models. Finally, it explains how the best performing model will be deployed to explore real data to classify Tweets not used to build the model.

3.1 Data Collection & Processing

Twitter's publicly available API was used with Python libraries to build a streaming Twitter data collector. Firstly, notable dates concerning important, publicised political discussions on the Grexit served as target dates for data collection. Tweets with the keywords 'greece' and 'euro' were collected over these time periods in 2015. The search term 'grexit' was omitted, because when the initial data was collected, it was not one of the most frequently noted words in tweets. Although it would produce more specific results, it may exclude opinions of those who are not familiar with the term (as it is more a term used in the financial sector). Moreover, 'grexit' tends to carry a certain connotation already, focusing only on Greece leaving the Eurozone, rather than the economic issues as a whole.

Following this, a second script was created to process the data, primarily using Python's NLTK [27]. Firstly, only tweets set to be in the English language were extracted. Next, URLs within the data were replaced with 'URL', in order to determine the frequency of link sharing, rather than the most popularly shared links. This is because this study aims to classify moral value judgements, rather than a network analysis of information. Finally, hexidecimal codes for "emojis" were also removed, leaving plain text for coding and analysis.

The script then generated frequency counts for the 100 most common words, "hashtags" and bi-grams. For this analysis, bi-grams are pairs of consecutive words. Frequency counts and bi-grams are useful to gain an overall picture of what people are talking about, enabling a quick overview of the main topics in the corpus. This allows for quick confirmation that the Tweets include concepts relating to MFT. From there, a list of common stop words was applied. Stop words contain the most common words in a language and corpus. Removal of these words is noted to yield much more accurate predictions in linguistic processing and classification [28]. Therefore, the most frequent words in the Tweets were also added to the stop word list, including 'URL', 'greece' and 'euro'.

After cleaning the data, the Tweets and bi-grams were manually coded. The codes were initially based on the moral foundations dictionary, where the words, their related words and their synonyms were used to guide classification. Beginning with a dictionary-based approach was useful in order to garner a more tangible picture of lexical indicators for each of the foundations. However, since liberty was not included, a list of synonyms for this foundation was created. Then, detailed descriptions of each of the foundations were used to better un-

derstand the nuances in each foundation, as outlined in the work of Graham et al. [10] and Haidt [9]. So the combination of specific, related words as well as detailed descriptions of the foundations were used to code the Tweets.

Feature selection is the process of selecting a subset of relevant features. For this classifier, additional features were based on the MFD, as well as the bi-grams produced by the second script. The most informative features were also extracted from the coded Tweets and added to the additional features. Through showing the most informative features, the classifier also highlights important keywords. Thus, unexpected mapping may occur when keywords emerge that are not included in the MFD, or if mapping appears between dictionary words and moral foundations that are different from the intended relations. A subset of these features, (the bi-grams) were coded by two individual coders, to measure agreement of classification among coders. This agreement will also be used to reflect on the accuracy performance of the classifier.

3.2 Learning

Supervised machine learning is used to classify Tweets. This method was chosen over unsupervised techniques, as the latter detects latent structures in text, which would not be useful for determining the specific values outlined in MFT. This section briefly outlines the principles behind Multinomial Naïve Bayes (NB) and Maximum Entropy (ME) classification.

These two classifiers were selected over other approaches such as Singular Value Decomposition (SVD), since SVD is noted to be more useful for large texts and can be used for feature extraction in combination with other learning algorithms [8]. Previous research using the MFD was conducted on long texts, therefore alternative methods should be looked into for shorter texts such as Tweets. Since Tweets are short, single-label output (one label per Tweet) was chosen over multi-output (multiple labels per Tweet).

Multinomial NB and ME algorithms were used to determine which machine learning approach produces the highest accuracy. Both are based on the application of Bayes' theorem: $\mathrm{P}(c|d) = \frac{P(d|c)P(c)}{Pd}$, with c the moral foundation class and d the Tweet (the document). They both apply Bayes rule to calculate the class of the Tweet, returning the one with the highest probability. They are both frequently used in solving classification problems, where in this case, each of the moral foundations is a class within the model.

Firstly, NB is noted to be a useful, scalable model for problems of classification, and is often used in spam filtering algorithms [29, 30]. ME is similarly popular, yet it is notably slower and is more useful for when contextual information is necessary for accuracy. This contextual information refers to the consideration of the relationship (correlation) between keywords, as opposed to estimating class membership from frequencies of individual keywords. For this study, both classifiers are interesting to examine as the accuracy of moral classification of tweets may be dependent on whether or not the data is examined contextually.

Despite their Bayesian origins, there are differences between the models. The differences lie in the probabilistic foundations of the NB and ME classifiers. NB operates under the assumption of conditional independence, meaning that it is assumed that there is no correlation between words. When the independence assumption is violated, it can lead to issues of double counting if certain words are highly correlated. On the other hand, ME theoretically posits that the probability distribution that is most representative of the data is the one with the largest entropy and takes into account the correlations between features. Therefore, ME is often used when the prior distributions of classes are unknown, and the conditional independence assumption is violated. Both classifiers return the class with the highest probability. Further information on the mathematical basis of each model can be found in Manning and Schütze's book on the Foundations of Statistical Language Processing [31].

Therefore, there are nuances in the statistical underpinnings of NB and ME classifiers. The most relevant key difference for this study relates to the independence of features, where NB assumes conditional independence and ME uses contextual information (such as N-grams) for classification. Both will be tested to determine which would be the best fit for classifying moral foundations in tweets.

3.3 Evaluation

To evaluate the data, Tweets were first labelled with the most appropriate moral foundation. Then, learning curves for the NB and ME algorithms were generated, extending the training data by 100 Tweets each run. Therefore, the training started at 100 tweets and end at 1,300 tweets. The remaining 700 coded tweets were used as testing data. The testing data remained the same for all runs.

Five-fold cross-validation was then used to provide an overall average accuracy of each model. This means that the data was split into 5 equal groups (N=400 per group). Each run used 400 Tweets in the test set, and 1600 Tweets in the training set. Each group of Tweets was the test set for one run. In the end, five accuracy results were obtained for each model, and the average of these results is reported as the final model accuracy.

Next, confusion matrices were used to determine the number of true positives (TP), false positives (FP) and false negatives (FN) for each foundation. A confusion matrix is used to visualise the performance of a supervised learning algorithm. From this table, the precision and recall of the classifier was calculated. Precision is measure of exactness of the classifier (where higher precision means less false positives), whereas recall measures the sensitivity of the classifier (high recall means less false negatives). The F-measure is the weighted harmonic mean of the two, which can be seen as a measure of overall classifier accuracy for each class. The F-measure is able to determine which foundation is most accurately classified by the algorithm. This is useful for future research, where accurate detection of certain foundations over others may be desired.

3.4 Deployment

The best performing algorithm was then trained with all labelled data. This trained classifier was then used to classify the rest of the unclassified tweets, generating insight into the dominant moral underpinnings of each data set. This enables analysis of the presence of the different moral foundations over time. It allows to discern if there are changes in moral concerns following the meetings regarding the Greek exit of the Eurozone.

4 Experiments & Results

This section discusses the results of the collected data, beginning with the baseline results. Then experiments using different training sets were conducted, first using raw labelled Tweets, followed by the supplementation of bi-grams and the MFD to the training data. Next, overall accuracy for each classifier is reported, as well as accuracy for each foundation class.

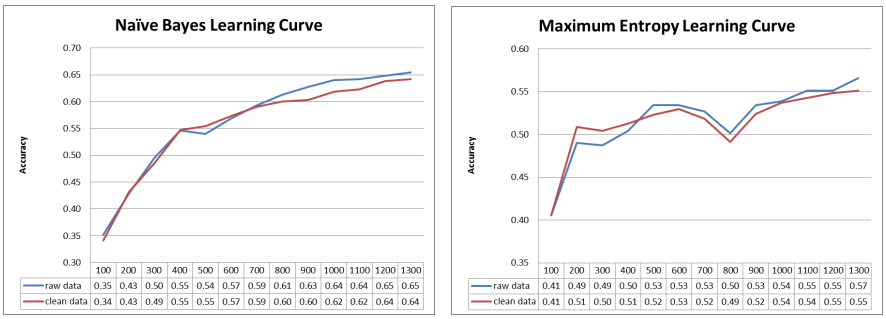
4.1 Baseline results

Data was collected from three different key points in time after Eurozone meetings, to garner initial reactions to the events. As such, there is a sampling of Tweets from the following time periods:

1. Feb 24 - March 3 (N = 7037): Eurozone Finance ministers agreed to extend the Greek bailout for another 4 months
2. April 28 - May 4 (N = 4856): Eurozone Finance ministers meet to discuss reform packages from Athens
3. May 11 - May 23 (N = 7066): Athens announces repayments to International Monetary Fund to avoid default

Each week of data collection yielded between 4-7 thousand data points (English Tweets), resulting in a total of 18,986 tweets collected. The duplicate entries were then removed (includes re-tweets and copy-pasting tweets from other statuses), leaving 8,292 unique tweets. From there, 2,000 Tweets were randomly selected and labelled as belonging to one moral foundation through manual coding.

In the labelled dataset, the most frequently labelled class occurred in 21.42% of cases, and thus the Zero Rules (ZeroR) majority vote benchmark accuracy is 21.42%. To determine the degree which coders could agree on moral classes, two coders labelled a dataset of bi-grams (N = 112) where coders agreed on 66% of the classifications. Even though coding bi-grams is different to coding full Tweets, it gives an indication of inter-coder agreement in classifying moral foundations. Therefore, any accuracy higher than the ZeroR value is noted to be an improvement of the classifier over random selection of the most frequently occurring class, and any accuracy around 66% would show that the classifier is as agreeable as human classification.



(a) NB Learning (b) ME Learning

Figure 1: Classifier accuracy by increase of training data

4.2 Experiment Set 1: Varying training data

Firstly, the data was split into training and test sets, with the testing set at 700 Tweets, and the classifier being trained in increments of 100 new Tweets, up until a maximum of 1300 Tweets, shown in figure 1. The classifiers were trained on conditions of: 1) raw data 2) data which had stop words removed (clean data) 3) raw data with the MFD, 4) clean data with the MFD and 5) raw data with labelled bi-grams as additional features.

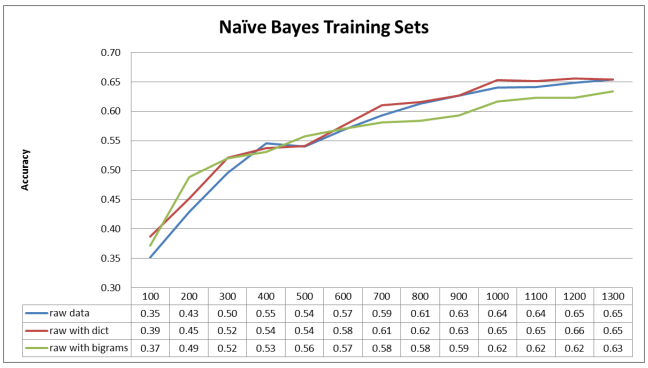
Interestingly, both classifiers performed slightly better without the removal of stop words (see Fig. 1). Hence there may be certain features which are removed that contribute to more accurate classification. NB showed higher overall accuracy (raw = 65%, clean = 64%) than ME (raw = 57%, clean = 55=%). Therefore, removing stop words did not increase classifier accuracy for either algorithm.

Since raw-data outperformed clean data, these training sets were appended with the MFD and labelled bi-grams in separate experiments (see Fig 2). Classifier accuracy shows that NB performance is not improved by the dictionary (both producing 65% accuracy). Moreover, NB performs worse than raw data when bi-grams are added, as accuracy drops to 63%. ME on the other hand, performs best with the addition of bi-grams. The features were also appended to clean data for both algorithms, but these are not reported here for the sake of brevity.

Figure 3 shows that over time, the learning curves of both classifiers flattens. It is therefore expected that additional training data will not improve classifier accuracy.

4.3 Experiment Set 2: Classifier Evaluation

Under all conditions, the NB classifier outperformed ME, shown in Fig 3. This result was confirmed by 5-fold cross-validation, where the mean accuracy for NB was 64.7% (SD = 0.03, p = .000) compared with the ME mean accuracy of 54.2% (SD = 0.02, p = .000). The difference in classifier accuracy is significant



	100	200	300	400	500	600	700	800	900	1000	1100	1200	1300
raw data	0.35	0.43	0.50	0.55	0.54	0.57	0.59	0.61	0.63	0.64	0.64	0.65	0.65
raw with dict	0.39	0.45	0.52	0.54	0.54	0.58	0.61	0.62	0.63	0.65	0.65	0.66	0.65
raw with bigrams	0.37	0.49	0.52	0.53	0.56	0.57	0.58	0.58	0.59	0.62	0.62	0.62	0.63

(a) NB learning conditions

	100	200	300	400	500	600	700	800	900	1000	1100	1200	1300
raw data	0.41	0.49	0.49	0.50	0.53	0.53	0.53	0.50	0.53	0.54	0.55	0.55	0.57
raw with dict	0.39	0.47	0.48	0.49	0.52	0.52	0.49	0.51	0.54	0.54	0.54	0.54	0.55
raw with bigrams	0.41	0.50	0.48	0.51	0.53	0.54	0.53	0.52	0.54	0.55	0.57	0.56	0.57

(b) ME learning conditions

Figure 2: Learning curves showing classifier performance with increasing training set size

(T = 13.9, p = .000). Overall, the NB classifier is 10 percentage points more accurate than ME in classifying moral foundations.

4.4 Experiment Set 3: Class Accuracy

All in all, the NB model produced the most accurate classifier, trained on raw data with the addition of the MFD to the feature set. This model was later used to classify each of the datasets which will be discussed further in Section 4.5. As the NB classifier was the best performing, it was examined further to determine which foundation could most often be accurately classified. This was confirmed by a confusion matrix (see Table 2). The matrix shows actual classifications in the rows, and predicted classifications in the columns. Therefore one can see which classes were correctly classified, as well as incorrect classifications as other classes. The confusion matrix in Table 2 shows that 'care' was most frequently classified correctly, followed by authority. Liberty was the least often correctly classified foundation. These experiments were also run for the ME classifier, but

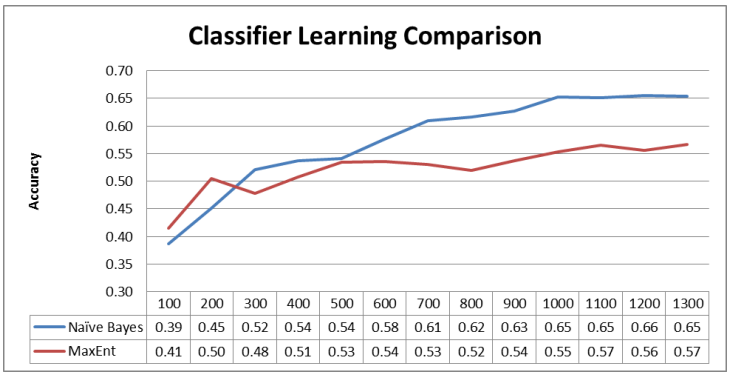


Figure 3: Overall classifier accuracy comparison

	authority	care	fairness	liberty	loyalty	sanctity
authority	< 87 >	13	2	.	9	4
care	6	< 108 >	7	5	13	7
fairness	8	25	< 61 >	7	23	6
liberty	4	12	4	< 38 >	10	2
loyalty	10	15	5	1	< 77 >	8
sanctity	10	23	3	.	14	< 73 >

Table 2: Naïve Bayes Confusion Matrix, comparing actual frequencies (rows) and predicted frequencies (columns)

again results are not reported here since it was the worse performing classifier.

The confusion matrix only shows the frequency of correct classifications, not in terms of relative accuracy. Therefore, from this data True Positives (TP), False Negatives(FN) and False Positives (FP) were used to calculate precision, recall and overall F-Measure (in Table 3). Authority was the most accurately classified (F = 0.73), followed by sanctity (F = 0.66) and care (F = 0.63). Fairness was the least accurate (F = 0.58).

Conclusively, the NB classifier outperformed the ME classifier and worked best being trained with raw data and the addition of the MFD. Adding coded bi-grams do to the feature space does not increase classifier accuracy. Thus, raw coded Tweets are sufficient for training a NB algorithm for moral foundation classification.

4.5 Application of classifier to Grexit Datasets

As mentioned in the section 4.1, there were 3 different time frames where Tweets were collected. The NB classifier was applied to each of these datasets, classifying each tweet individually (N = 16,986). These Tweets do not include those which were used in the training and testing set of the classifier. It can be seen

	TP	FN	FP	Precision	Recall	F-Measure
authority	87	28	38	0.696	0.757	0.725
care	108	38	88	0.551	0.74	0.632
fairness	61	69	21	0.744	0.469	0.575
liberty	38	32	13	0.745	0.543	0.628
loyalty	77	39	69	0.527	0.664	0.589
sanctity	73	27	27	0.73	0.593	0.655
Total	444	256	256			

Table 3: Naïve Bayes accuracy for each class

in Figure 4 that the most frequent Tweets were classified by the NB algorithm in the 'care' foundation. Thus, over the course of 2015, individuals on Twitter showed 'care' as the primary moral concern in the Grexit debate, authority as the second, and loyalty as the third.

However, over time, the predominant moral underpinning of the rhetoric can change. Indeed, Figure 5a shows that in the first and third time periods, 'care' was the most common concern, whereas in the second time period, 'authority' dominated the discussion overall. In all time periods, 'liberty' was the least discussed foundation, especially in Dataset 2, where the foundation barely emerged (see Figure 5c Thus, in light of the hypothesis that 'liberty' is a necessary foundation for this research, application of the classifier shows that people on Twitter are not primarily concerned with liberty or oppression of any party in this debate.

The classification of Tweets per day is shown in figure 5. Datasets are numbered 1, 2 and 3 respectively, and the individual days of collection are noted as $1.1, 1.2 \cdots X.n$. For each data set, it appears that overall discussion would begin frequently directly following the meetings, decrease for a few days, and increase once more 3 - 4 days after.

In the first dataset, seen in Figure 5b initial public reaction was rooted in 'authority' and 'loyalty', but 'care' quickly rose the following day, and was the dominant foundation of discussion 3 days following the meeting. Towards the end of the week, 'loyalty' and 'fairness' was more frequently driving the discussion.

The foundation of 'loyalty' was also a major starting point of discussion on the first day of Dataset 2, alongside 'authority'. Figure 5c clearly shows that the early discussion was dominated by 'authority' foundations, and in the later half people were far more concerned about 'care'. Overall however, it is shown in Figure 5a that the second dataset was mainly with concerned with 'authority'.

The third Dataset spans over almost 2 weeks, and therefore there are multimodal points of interest. For instance, Figure 5d shows that for the first week, 'care' was the foundation of highest concern, yet 'sanctity' become of importance following the first week. Conversely, while 'authority' was an important point overall, this dataset shows a trend of decreasing emphasis on 'authority'.

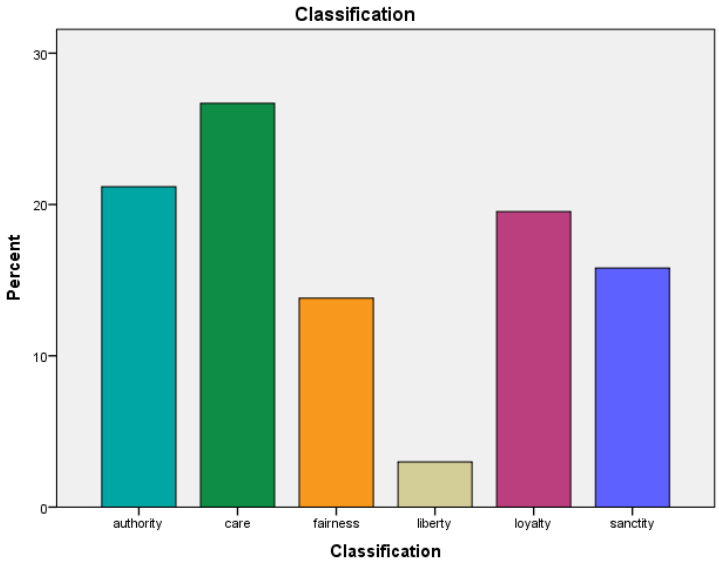


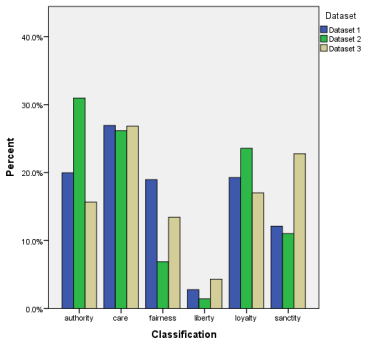
Figure 4: Classification of all Tweets

Moreover, this was the only dataset where liberty started to be a more frequent point of discussion, especially after the 9th day of collection.

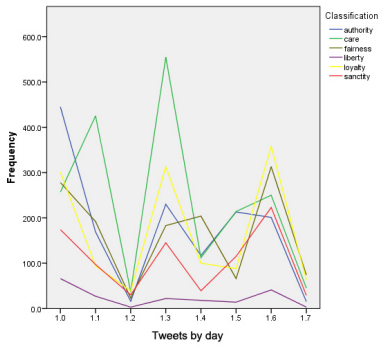
Interestingly, 'care' being the most frequently classified foundation shows that public discussion is not in line with the analysts moral view of the situation, which is more related to fairness and loyalty. From this, it can be inferred that English-speaking people on Twitter are perhaps more concerned about protecting others, or the identifications of victims in this situation, rather than any other morally driven opinion. Conversely, 'care' could be seen from different ideals - caring for Greece and keeping them in the Eurozone by bailing them out, or care for the European Union by removing nations with bad economic performance or policies. Since 'care' and 'authority' were the most present foundations, one can ascertain that there is some discrepancy between public value judgements and that which are posited by authority figures.

5 Discussion & Recommendations

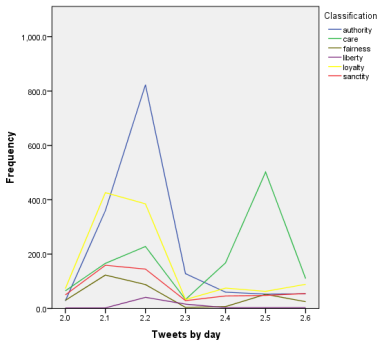
The results show that the NB classifier is a good starting point for attributing moral foundations to Tweets. The F-Measures for the three most accurately classified foundations (care, authority and sanctity) are somewhat in line with previous research regarding agreement of virtuous terms for these foundations [16]. Thus, NB is a useful machine learning tool for classifying moral foundations on Twitter, especially those which have virtues that extend beyond cultural boundaries.



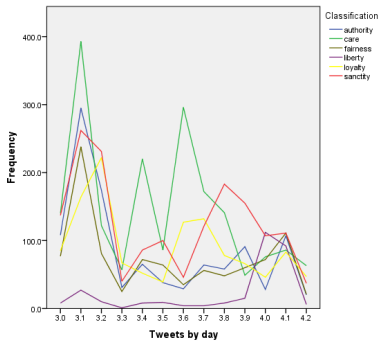
(a) Classification of Tweets over Time



(b) Dataset 1: February 24 - March 3, 2015



(c) Dataset 2: April 28 - May 4, 2015



(d) Dataset 3: May 11 - May 23, 2015

Figure 5: Applied classifier output

The performance of the NB algorithm for correctly classifying the 'liberty' foundation was mediocre. It was not the most accurately or frequently detected class. Conversely, 'liberty' was the least frequently classified foundation and was least discussed in all of the time periods. Despite the algorithm showing some accuracy in correctly classifying the foundation, results were not as clear-cut as expected. Therefore, until further research for lexical indicators of 'liberty' is conducted, it is recommended that this foundation be excluded from future research in this context.

Results also showed that addition of features from the MFD did not improve classifier accuracy. Therefore, the usefulness of the MFD in a frequency based classification approach is called into question. If the use of this dictionary is desired in future research, improvements to the dictionary should be made by adding words belonging to the most informative features identified in the NB algorithm. However, the efforts in improvement of the MFD may only have marginal implications for classifier accuracy, and it is recommended to drop it all together as these dictionaries are costly to build and maintain.

Despite promising outcomes, there are several limitations which must be con-

sidered. Firstly, although the tweets are able to be classified into one foundation, multi-label output may be more appropriate when studying moral foundations. This is because human moral expression is exceedingly complex and people themselves can disagree on the related foundation, especially without elaborate contextual information. Thus, single-label classification does not capture the subtleties in morality that is displayed in human communication.

Inherently, there are also differing perspectives on predominant moral values. Coding Tweets using MFT was cumbersome and difficult, as each foundation needed to be carefully operationalized by those with a sound understanding of the theory. On a positive note, the learning curves show that coding more than 2,000 Tweets for training and testing of a classifier will not improve accuracy. But if one wishes to apply this method to other datasets, it would not be possible to use a mechanical Turk or outsource the job to those who are not familiar with MFT. This type of research calls for a truly interdisciplinary approach, where the right expertise is needed from all fields involved. Yet, expensive expert resources are better spent on labelling Tweets as opposed to building dictionaries.

There are also ethical concerns about using open data from individuals and making the results of this data open. In other words, there is a moral consideration of mining opinions of individuals [32]. Mining moral opinions can lead to building personality profiles and knowing intimate details about individuals that can be used in different targeting manners. This concern was addressed, as all Tweets were anonymized and can not be tied to individuals without additional web search efforts.

Although the ME classifier performed worse, it is recommended to apply different methods of feature selection for further testing, since it was best preforming with the addition of bi-grams. Therefore, the algorithm may perform more accurately and faster with a reduced feature space that can be created by frequently used word collocations extracted from the corpus. This would also enable quicker and more efficient manual coding. Although, the price of this feature selection would be that certain nuances in the Tweets may be missed, resulting in misclassification.

Conversely, multi-label output is another method that may be more fitting to classifying moral foundations. Even with short Tweets, there can be several different moral foundations present. If continuing with the NB approach, probabilities for each class per Tweet can be produced. However, this would greatly add to the time needed for data analysis. Through using multi-label classification, one would need to sift through vast amounts of results to draw conclusions. Semi-supervised learning may also be an interesting approach to use together with multi-label output, as it focuses on minimizing labelling effort.

Lastly, it was considered that all Tweets may not necessarily relate to a moral foundation. However, manually labelling Tweets made it clear that moral concerns are present in the data. In other words, morality is a reason to discuss the topic in the first place. All in all, there are advantages and limitations to using an NB classifier to study moral foundations. Yet, for specific controversial issues, classification of all Tweets into moral foundations is possible and appropriate.

6 Conclusion

This study presents an experiment to see if we can better understand moral value expression on Twitter through using machine learning methods. The research question aimed to examine the extent supervised machine learning techniques can learn to detect moral foundations in Twitter communication. The best performing algorithm could make moral classifications on a similar level to human coders.

Contrary to the earlier hypothesis, liberty was not a frequently classified foundation in any time period. It may not be appropriate to include the class in future research, as it did not seem to appear often in the discussion, or was cannibalised by the 'authority' foundation, which has much overlap. Perhaps only with more concrete definitions it may be worthwhile to continue exploration of this additional foundation.

In the future, this classifier can give insights into the most informative features for each class. With some insight to the key predictors, feature sets could be optimised for each class. With this optimization, the classifier could essentially be applied to any controversial topic discussed on social networks.

On the whole, the NB classifier was 10 percentage points more accurate than the ME classifier, and the accuracy is comparable to the agreement of moral classification between humans(64.7% compared with 66%, respectively). Moreover, it is roughly 3x more accurate than the baseline ZeroR measure of 21.4%. Hence, using an NB classifier is a good starting point for single-label classification of moral foundations. At this point, it is difficult to compare with other moral foundation classification research, as thus far none have used the NB or ME classifiers. Conclusively, NB is a machine learning algorithm which can classify Tweets by their moral foundations as well as human coders.

References

[1] K. Kinder-Kurlanda and K. Weller, "i always feel it must be great to be a hacker!: the role of interdisciplinary work in social media research," in *Proceedings of the 2014 ACM conference on Web science*. ACM, 2014, pp. 91–98.

[2] Z. Papacharissi, "The virtual sphere the internet as a public sphere," *New media & society*, vol. 4, no. 1, pp. 9–27, 2002.

[3] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[4] D. Freelon, "On the interpretation of digital trace data in communication and social computing research," *Journal of Broadcasting & Electronic Media*, vol. 58, no. 1, pp. 59–75, 2014.

[5] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: mapping the spread of astroturf in microblog streams," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 249–252.

[6] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011, pp. 30–38.

[8] E. Sagi and M. Dehghani, "Measuring moral rhetoric in text," *Social Science Computer Review*, vol. 32, no. 2, pp. 132–144, 2014.

[9] J. Haidt, *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[10] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," *Advances in Experimental Social Psychology, Forthcoming*, 2012.

[11] P. De Grauwe, "The eurozone as a morality play," *Intereconomics*, vol. 46, no. 5, pp. 230–231, 2011.

[12] J. Bond, "It ain't over till the fat lady sings," *Sens-Public*, 2012.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[14] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[15] S. Clifford and J. Jerit, "How words do the work of politics: Moral foundations theory and the debate over stem cell research," *The Journal of Politics*, vol. 75, no. 03, pp. 659–671, 2013.

[16] M. Dehghani, K. Sagae, S. Sachdeva, and J. Gratch, "Linguistic analysis of the debate over the construction of theground zero mosque'," *Journal of Information Technology & Politics*, vol. 11, pp. 1–14, 2014.

[17] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations." *Journal of personality and social psychology*, vol. 96, no. 5, p. 1029, 2009.

[18] J. Haidt and C. Joseph, "Intuitive ethics: How innately prepared intuitions generate culturally variable virtues," *Daedalus*, vol. 133, no. 4, pp. 55–66, 2004.

[19] D. Gros, "Grexit 2015: A primer. ceps commentary, 22 january 2015," 2015.

[20] A. Lazarou *et al.*, "Greece: The many faces of yanis varoufakis," 2015.

[21] A. Randazzo and J. Haidt, "The moral narratives of economists," *Econ Journal Watch*, vol. 12, no. 1, p. 49, 2015.

[22] Y. Varoufakis, "Egalitarianism's latest foe: a critical review of thomas piketty's capital in the twenty-frist century," *real-world economics review*, p. 18, 2014.

[23] C. B. May, "Greek euro-crisis: Consequences of a "grexit"."

[24] C. Mascaro, A. Black, and S. Goggins, "Tweet recall: examining real-time civic discourse on twitter," in *Proceedings of the 17th ACM international conference on Supporting group work*. ACM, 2012, pp. 307–308.

[25] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 1320–1326.

[26] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, "Mining contrastive opinions on political texts using cross-perspective topic model," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 63–72.

[27] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009.

[28] H. Saif, M. Fernández, and H. Alani, "Automatic stopword generation using contextual semantics for sentiment analysis of twitter," in *CEUR Workshop Proceedings*, vol. 1272, 2014.

[29] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161–168.

[30] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[31] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[32] M. Christen, M. Alfano, E. Bangerter, and D. Lapsley, "Ethical issues of 'morality mining'."