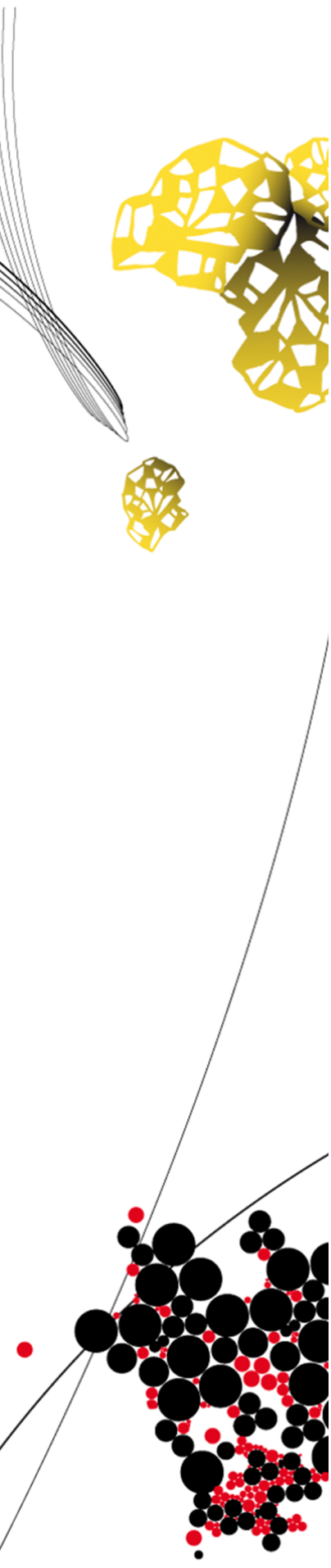


June, 2024

Master Thesis
Interaction Technology

Implications of Detection of Moral Foundations in Written Text

Thijs de Kleijn
s1485830

Supervisors
dr. Lorenzo Gatti
dr. Estefanía Talavera Martinez
dr. Mariët Theune

UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science
Human Media Interaction group

ABSTRACT

The Moral Foundation Theory is a way of classifying intuitive moral behaviour. The intuitive and fundamental nature of the Moral Foundation Theory theorizes that people will fall back on the morals they find important before logical reasoning. These morals also affect language used and understanding their presence can help with better understanding the underlying message. Detecting the presence of moral foundations in a piece of text can be interesting for the psychology domain, but such a task requires knowledge in the natural language processing domain. This thesis tries to bridge the gap from natural language processing to psychology and elaborate on steps taken for granted within the natural language processing community.

In particular, this research compares and analyses text representations methods and classification algorithms, and tests their suitability for cross data set classification with the available data sets. We use two large data sets with annotations reflecting the Moral Foundation Theory: the Moral Foundation Twitter Corpus and the Moral Foundation Reddit Corpus. Based on majority voting, a single label is selected from all annotations for each post.

All experiments are performed in two variations of classification: moral against non-moral and morals-only. For text representation methods, we compare a general Word2Vec embedding, GloVe's pre-trained Twitter-200 model, against a dedicated dictionary based on the Moral Foundation Theory, the extended Moral Foundation Dictionary. The different classification algorithms are Logistic Regression, Support Vector Machines and distilBERT. Lastly, we test the performance for cross data set classification.

The results show that moral against non-moral classification is successful regardless of text representation or classification methods, whereas morals-only classification is only successful with GloVe's representation. Comparing the classification algorithms, distilBERT generally has better performance, but does not strictly outclass Logistic Regression or Support Vector Machines.

Unfortunately, cross data set classification is not successful with the data sets at hand.

Future work should consider improving on text embedding techniques, returning more classification outputs to cover the ambiguous nature of the Moral Foundation Theory, and aligning the theme of the training data set to the testing data set.

TABLE OF CONTENTS

LIST OF ACRONYMS

ALM	All Lives Matter
BERT	Bidirectional Encoder Representations from Transformers
BLM	Black Lives Matter
eMFD	extended Moral Foundation Dictionary
GloVe	Globalized Vectors
LIWC	Linguistic Inquiry and Word Count
LogReg	Logistic Regression
MFD	Moral Foundation Dictionary
MFRC	Moral Foundation Reddit Corpus
MFT	Moral Foundation Theory
MFTC	Moral Foundation Twitter Corpus
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PABAK	Prevalence
RNN	Recurrent Neural Network
SVM	Support Vector Machine
W2V	Word-to-Vector
Word2Vec	Word-to-Vector

1 INTRODUCTION

The Moral Foundation Theory (MFT) of Haidt and Joseph [1] aims at generalising the underlying elements that help humans with (moral) decision making. Haidt and Graham [2] claim that these elements are universally applicable to all humans regardless of background or culture. Morality in general allows humans to make quick decisions between good and bad choices. In a more primitive sense, this means choosing the option which is best suited for survival. As such, intuition activates first and foremost before reasoning comes into play. These principles make up the fundamentals of moral decision making and are therefore called the foundations. Haidt and Joseph [2] noted five distinct foundations. To reflect the good and bad choices of morality, each foundation is split into a positive aspect called virtue and a negative aspect called vice. The foundations are Care & Harm, Fairness & Cheating, Loyalty & Betrayal, Authority & Subversion, and Purity & Degradation. The foundations are also shown in Table 1.1 with a short example reflecting the essence of each foundation.

Moral decisions appear to be quite complex and as such, most decisions can be influenced by multiple foundations at a time. The intensity and the combinations of foundations are both culturally and personally dependent [2]. While the Moral Foundation Theory is just one approach within the psychological domain, its framework helps with identifying typical moral issues that can arise when faced with decision making.

Graham and Haidt [3] identified that certain use of language can be associated with specific moral foundations. Understanding this relation between language usage and moral concern can help with gaining a better understanding of the implications of a message by highlighting the underlying moral foundations. This could help with discussion partners getting a better understanding of each other or an e-health chatbot system more successfully conveying a message related to behaviour change. At a larger scale, it can prove beneficial for polling (public) community opinions or studying the (public) communication of a political entity.

The existence of the Moral Foundation Theory implies that some psychology domains have an interest in classifying moral foundations and even linking them to text. The field of Natural Language Processing (NLP) has ample and elaborate tools available for extracting information from text. Unfortunately, psychology and NLP practicioners typically have limited overlap. As such, this thesis tries to bridge the gap from NLP to psychology and investigate some of the implications of tools used for classifying moral foundations from text, in addition to elaborating some steps taken for granted within the NLP community.

Table 1.1: Foundations of the Moral Foundation Theory listed as virtues and their vices supported by an example based on the definitions created by Haidt and Joseph [2].

Virtue	Vice	Example
Care	Harm	Protecting and caring for children.
Fairness	Cheating	Equal treatment of actions to benefit each of the involved individuals.
Authority	Subversion	Create beneficial relationships within hierarchies. Being dutiful towards your leader warrants protection and a stable society.
Loyalty	Betrayal	Forming small groups for survival, cooperation, and friendship.
Purity	Degradation	Avoiding contaminants to maintain a sufficient degree of health, both physical and mental.

1.1 Main Research Question

To what extent can Natural Language Processing techniques identify Moral Foundations in text?

1.1.1 Sub Research Questions

This research will consist of a comparative analysis between the elements mentioned below. As such, each important aspect of Natural Language Processing can be investigated for its contribution to detection of Moral Foundations and give insight into improvements that can be made.

- To what extent do different text representations influence the results?
- What are the implications of using deep learning or machine learning techniques and how do they compare in performance and usability?
- How well do the available corpora lend themselves for cross data set training and testing?

1.2 Overview

This thesis focusses on detection of moral foundations using NLP approaches, as described by the following set up: In Chapter 2 Background where we investigate similar research to obtain techniques and data sets for performing classification of moral foundations in text. The obtained data sets are cleaned up and sorted out in Chapter 3 Data Preparation. The text data is then vectorized before Chapter 4 Method describes the setup for all experiments required to reach the necessary results, which in turn are reported in Chapter 5 Results. The results are supplemented with baseline thresholds and results of similar experiments from literature.

Chapter 6 Discussion puts the results in perspective to the research questions. Additionally, this chapter elaborates on imitating factors and notions for future work encountered during this research. We end this thesis by summarising the answers to the research questions and the main contributions of this research in Chapter 7 Conclusion.

2 BACKGROUND INFORMATION

In this chapter we provide a more detailed overview of the Moral Foundation Theory. In addition, we introduce the data sets and the techniques from related works that have used the Moral Foundation Theory in natural language processing tasks.

2.1 Moral Foundation Theory

The moral foundation theory (MFT) is already briefly mentioned in the Introduction, Chapter 1, as a reflection of a certain primal intuition. Designed by Haidt and Joseph [1], the MFT attempts to classify intuitive moral behaviour. These intuitions dictate feelings and potentially actions before any logical reasoning comes into play. As such, this is called intuitive ethics and serves as a form of emergency response. Haidt and Graham [1] explored and investigated intuitive patterns into five main categories. Each category is called a 'foundation' or 'dimension' and has a positive and a negative side, called 'virtue' and 'vice' respectively[1].

The first foundation is *Care*, with *Harm* representing the vice. Care is characterised as arguably the most primal foundation [2]. It relates to protecting someone who is unable to protect themselves. This virtue is typically exerted in acts of kindness and compassion and the vice as cruelty and aggression.

Fairness and *Cheating* deal with equal or discriminatory treatments. This foundation is paramount for raising judicial institutions and upholding the law. In weaker term, Fairness is also reflected in writing reviews. Atari et al. [4] propose splitting Fairness into Equality and Proportionality, to reflect equal opportunity and equal results respectively.

The innate human drive to form social connections is reflected in the foundation of *Loyalty* and *Betrayal*. Belonging to a tribe or supporting a sports team, as well as traits such as heroism and patriotism, are common themes in this foundation. Conflicts over religion or other beliefs usually originate from different conceptions of this foundation.

Authority and *Subversion* represent the evolutionary history of having hierarchies in social interactions. Having someone in charge is probably preferred over anarchy, but this may not be the case if the leader is a tyrant or dictator.

The final foundation is *Purity* and *Sanctity*. This foundation is concerned with avoiding infectious or otherwise contaminating substances to our bodies. This is typically expressed in traditional beliefs, religion, and general sinful behaviour on both social and health topics.

Aside from these five main foundations, Iyer et al. [5] identified another unique foundation; *Liberty* and *Oppression*. This dimension covers the freedom of choice and the absence of personal freedom. As such, it highlights one of the main traits of libertarian minded people.

The Moral Foundation Theory is not a flawless classification of intuitive moral behaviour. This is also proven by the additions of Atari et al. [4] and Iyer et al. [5]. Moral decisions are complex and a single decision can be influenced by multiple foundations and in some cases the presence of one foundation could negatively impact the presence of another founda-

[1] The foundations are mainly named using just their virtue. E.g. The foundation of Care covers the whole range from Care to Harm

tion. Even more so, the intensity and the combination of foundations are both culturally and personally dependent. For example, United States Democrats generally value Care and Fairness more than Loyalty, Authority and Purity, which in turn are more valued by United States Republicans [2]. Haidt and Joseph [1] also highlight the importance of understanding the moral principles and ethics that shape the thoughts of people in a discussion. As such, the main purpose of the MFT is to help categorising intuitive moral and ethical behaviour such that other research has some guidelines to follow.

2.2 Text representation

Text on its own is not really usable for computations. Aside from comparisons, raw text does not hold the same value for computers as we as humans associate to it. Text representation methods transform raw text into usable data for algorithms.

2.2.1 Embedding techniques

We use text representations to transform text into usable entries for further processing and calculations. Different representation methods bring different positive and negative aspects to the table. We explore text representation methods used by similar research, focussed on classifying moral foundation, or otherwise involving them in natural language processing tasks.

Linguistic Inquiry and Word Count (LIWC) [6] is based on the idea that psycholinguistic information provided better results in an affective NLP task. LIWC makes use of feature vectors with 64 categories covering themes such as well-being, culture, social behaviours, and cognition. For each word in a piece of text, the feature vector is updated for all categories attributed to that word. The resulting vector is normalised with respect to the length of a piece of text. To ensure a wider usability, the entries within the LIWC are stems so that variations of the same word are not ignored. For example, the stem ”ador*” will cover words like ”adore”, ”adoration”, and ”adoringly”.

Global Vectors (GloVe) [7] is an unsupervised learning algorithm for creating word vector representation and is supplemented with pre-trained word vector models, based on Mikolov’s Word2Vec approach [8]. GloVe highlights three main aspects to ensure that semantic information is maintained, namely word analogy, word similarity, and named entity recognition. Word analogy ensures that the relation between *man* and *woman* is similar to the relation between *king* and *queen*. Word similarity clusters related words together such that the vectors for related words have short Euclidean distances. These clusters not only includes words that are written similarly, such as *frog* and *frogs*, but also includes *toads* and the names of different species of frogs. Lastly, named entity recognition helps with identifying (important) persons, locations, and organizations. Everything combined, the heavy focus on semantics ensures that GloVe models retain all the necessary information to highlight the message portrayed in a piece of text.

2.2.2 Dictionaries

Dictionaries are a specific text representations for detecting moral foundations. Dictionaries help to assign a certain value to words that are related to the MFT. This allows us to make statistical and numerical comparisons on the morality that can be obtained from a text. Thus creating insights in the underlying morals of a text.

The initial Moral Foundation Dictionary (MFD), summarised in Table 2.1, is created by Graham and Haidt [3]. This dictionary contains a limited set of annotated words and stems. Each dictionary entry was manually selected and labeled. These entries were attributed a

single label according to the five dimension of the MFT and their virtues and vices. In the case of more ambiguous entries, the attributed label became ”General Morality”. This results in a total of eleven unique labels that can be distributed. While most words or stems are only given a single label, some have two or even three unique labels to cover its versatile nature of the Moral Foundation Theory.

The Moral Foundaiton Dictionary 2.0 (MFD2.0, [9]), summarised in Table 2.1, extends the original MFD regarding the amount of words and removes stems in the process. The resulting dictionary includes proper words as well as short phrases, such as "*us against them*". Similar to the initial MFD, the MFD2.0 starts with a set of manually selected words that appear relevant for each foundation. This averages to 210 words per foundation. From hereon out, eight seed words have been selected for each foundation, four for the virtue and four for the vice. These seed words have been tested using Word2Vec [8] representation schemes [9]. Based on a threshold, all potential candidates are filtered for their relevance with respect to the initial seed words. As such, only words that are the most prototypic for their respective foundation are kept in order to not diminish the dictionaries' validity.

Hopp et al. [10] state that the initial MFD, and therefore also the MFD2.0, are a straightforward way of dealing with automatic moral information extraction. However, major shortcomings include the limited selection of words per foundation, which are manually selected by a small group of experts. In addition, words are labeled according to a single foundation. This leads to conflict with words that could be valuable in the context of multiple foundations. As such, the extended Moral Foundation Dictionary (eMFD, [10], summarised in Table 2.1) takes a new approach to creating a dictionary in order to provide a more detailed insight in the moral load of words. First of all, Hopp et al. [10] managed to increase the dictionary size to 3270, an increase of roughly 30% when compared to the MFD2.0. In addition, all entries are provided with both probability and sentiment scores, scaling between -1 and 1. The probability scores represent the probability of encountering the given word in the corpus. The sentiments scores indicate how positive or negative a certain entry is rated. With +1 indicating a perfect virtue and -1 indicating a perfect vice. As such, this score represents the moral valence of an entry.

2.2.3 Chosen Text Representation

Pavan et al. [13] have used LIWC as a word representation in order to detect the presence of moral foundations and have discovered that the LIWC word representation method does not provide accurate results.

Out of the remaining materials, we select the extended Moral Foundation Dictionary [10] for its direct relation to the Moral Foundation Theory in addition to being both larger and more specified than its earlier iterations. Lastly, GloVe's Word2Vec model [7] is chosen as a general text representation method and its ability to retain semantic information. This allows to test a general model against a dedicated model.

2.3 Classification Techniques

After applying text representation methods, we can look into classification methods and similar to other classification tasks in NLP, there are various ways to approach this. From similar research focussed on classifying moral foundations in text, [13–18], we identify three popular approaches for such a task: classic methods, Long Short-Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT, [19]). Classic methods refer to statistical models such as Naive Bayes, logistic regression, k-nearest

Table 2.1: List of MFT dictionaries, including a short description, size, and how the data is labelled.

Name	Description	Database Size
Moral Foundations Dictionary [3, 11]	Words and word stems are attributed one or more labels indicating the appropriate foundations	324 words and word stems. 11 tags: 5 virtues, 5 vices, 1 general morality
Moral Foundations Dictionary 2.0 [9]	Complete words or phrases, e.g. "*us against them*". Some words occur multiple times under different tags.	2103 words. 10 tags : 5 virtues and 5 vices
Extended Moral Foundations Dictionary [10]	Probability of appearing with the corpus and a rated value for each of the foundations.	3270 words, all labeled with 10 elements: 5 for probabilities for each foundation and 5 for sentiment for each foundation.
Moral Strength [12]	Values scaled between 1 (absolute vice) and 9 (absolute virtue).	996 words. 5 tags: Authority, Care, Fairness, Loyalty, Purity

neighbours, and support vector machines (SVM). These models are generally fast to train and easy to interpret.

Long Short-Term Memory is a recurrent neural network (RNN) aimed at retaining short-term information over a large amount of steps. As such, it can store long term dependencies so grammatical structures can be maintained making it a solid method for document-level classification tasks.

BERT is a high-end approach for classifications steps based on neural networks. BERT processes words in-sequence through multiple layers of encoding, each layer capturing more sequentially dependent information. Unlike LSTM, BERT does not use recurrent units and is therefore faster than LSTM.

2.4 Data sets

Using exclusively unsupervised learning methods risks obtaining results that are not in line with the Moral Foundation Theory. While this can still provide interesting results, supervised learning guarantees that we make predictions using the dimensions found in the Moral Foundation Theory. Fortunately, two big labelled data sets exist. The first containing Moral Foundation labels is the Moral Foundations Twitter Corpus (MFTC) [16][2]. To create this dataset Hoover et al. [16] collected 35,108 tweets and hand annotated them according to the virtues and the vices of the dimensions of the Moral Foundation Theory or non-morality. The selected tweets fall into one of the following seven categories, with their distribution specified in Table 2.3: All Lives Matter (ALM) movement, Black Lives Matter (BLM) movement, 2016 U.S. Presidential Election, Baltimore protests following the death of

[2] During this research, Twitter has officially been renamed to X since July 2023 [20]. This thesis uses the name Twitter in favour of X to align itself with similar research and important materials that are cited, referenced, and used.

Freddie Gray, #MeToo movement, Hurricane Sandy aftermath, and Davidson Hate Speech Corpus [21]. These categories were selected by Hoover et al. [16] based on their relevance to current problems in the social sciences and their high likeliness of containing a variety of moral concerns. Each tweet was manually reviewed and annotated by at least 3 annotators out of a set of 13 total trained annotators. Each annotators could label each tweet with multiple labels.

One of the main limitations of the MFTC is the use of tweets. Almost all tweets are short sentences up to 280 characters maximum, with most tweets averaging out to be only 34 characters [22]. Both the character limit and the proven average severely hamper the ability to pose nuanced statements or context.

Trager et al. [15] agree with the usefulness of the MFTC, but argue that its limitations are too hampering for proper research using NLP and the MFT. There are two main issues: the limited allowed characters lead to shorter messages and therefore less nuance surrounding the posted statement. Using posts from Reddit, Trager et al. hope to solve the issues that occur when using tweets, since Reddit allows for significantly larger posts. The organizational structure of Reddit with subreddits allows for easy access to on-topic posts. Reddit users are also provided an extra layer of anonimity, as opposed to e.g. Facebook or in some cases Twitter. This anonimity can help with more openly speaking one's mind, thus leading to better expression of opinions and other moral language. Trager et al. [15] have gathered 16,123 Reddit posts from 12 different subreddits, specified in Table 2.4, each selected for high-moral loading and the exclusion of bot posts indicated by "I am a bot" at the end of a post. Each post is annotated by a minimum of 3 and a maximum of 5 annotators out of a total set of 23 trained annotators. Each annotator labeled their sample of Reddit posts independent of other annotators and are allowed to assign multiple labels. There are 8 unique labels that could be attributed to a post: Care, Equality, Proportionality, Loyalty, Authority, Purity, Thin Morality, and Non-moral. In addition to attributing a label regarding morality, the annotators also included a notion of confidence: Confident, Somewhat Confident, and Not Confident. This helps with grading the subjectivity levels of the annotation.

Both the Twitter and the Reddit corpus are explained in more detail in Chapter 3 as well as preparing the provided data for further experiments.

Table 2.2: Labeled Corpora and their specifications

Name	Labels	Size
Moral Foundation Twitter Corpus [16]	Non-moral and all 5 dimensions and their vices	35,075 tweets from 7 different domains, see Table 2.3
Moral Foundation Reddit Corpus [15]	non- and thin-morality and 6 foundations	16,123 posts from 12 different domains, see Table 2.4

Table 2.3: Moral Foundation Twitter Corpus Twitter hashtags

Topic	Abbreviation	Selection Criteria	Database Size
All Lives Matter	ALM	#AllLivesMatter, #BlueLivesMatter	4,424
Baltimore Riots	Baltimore	All tweets from cities with Freddie Gray protests	5,593
Black Lives Matter	BLM	#BLM, #BlackLivesMatter	5,257
2016 U.S. Presidential Election	Election	Followers of the presidential candidates and official news outlets	5,358
Davidson Hate Speech Corpus	Davidson	Random sample from Davidson's Hate Speech corpus [21]	4,961
Hurricane Sandy aftermath	Sandy	#HurricaneSandy, #Sandy	4,591
#MeToo	Metoo	Subset of tweets mentioning user associated with allegations of sexual misconduct	4,891

Table 2.4: Moral Foundation Reddit Corpus specification of subreddits

Subreddit	notes	Database Size
r/AmItheAsshole		1339
r/Conservative	French politics	144
r/Conservative	U.S. Politics	1776
r/antiwork		1771
r/confession		1331
r/europe		2647
r/geopolitics		113
r/neoliberal		1673
r/nostalgia		1342
r/politics		1768
r/relationship_advice		1353
r/worldnews		2564

3 DATA PREPARATION

In this chapter, we retrieve the Moral Foundation Twitter Corpus and the Moral Foundation Reddit Corpus, clean them up and assign labels to each post. After this, all posts are subjected to text embedding to create vectors ready for further experiments.

3.1 Dataset Acquisition

3.1.1 Moral Foundation Twitter Corpus

The Moral Foundation Twitter Corpus (MFTC) [16] should contain 35,108 Twitter posts, with 11 possible labels. The MFTC only included reference codes to the posts and not the posts themselves. To get the posts themselves, we have to hydrate Twitter which means replacing the reference codes with the actual content of the referenced posts. The hydrating process involves scraping Twitter and using the reference code to obtain the associated post. During hydrating, 332 entries are either empty or not available for retrieval, these entries are immediately discarded, resulting in 34,776 total retrieved posts.

Each tweet was annotated by a minimum of 3 annotators and each annotator could assign any number of labels. All annotations are included in the MFTC, without reporting the most appropriate label. To obtain the most suitable label, all annotations are condensed into a single list and the label that occurs the most often is selected. In some cases a tie occurs, either due to multiple labels occurring equally as often or due to all annotations being completely unique. In this case, the label that appears first is automatically chosen. The resulting Twitter posts and their associated labels are structured like shown in Table 3.1 for easy access and processing later on. Finally, the resulting label distribution, shown in Table 3.2, differs from the distribution reported by Hoover et al. [16].

Table 3.1: Moral Foundation Twitter Corpus after hydrating and assigning labels.

	text	dimension	label
0	@fergusonoctober @FOX2now #AllLivesMat...	CareHarm	care
1	Wholeheartedly support these protests ...	AuthoritySubversion	subversion
2	This Sandra Bland situation man no disr...	FairnessCheating	cheating
3	Commitment to peace, healing and loving...	CareHarm	care
4	Injustice for one is an injustice for all #All...	FairnessCheating	cheating
5	This is what compassion looks like! #vegan...	CareHarm	care
6	@CNNPolitics @IngrahamAngle @phucbho ...	CareHarm	harm
7	Black Twitter when they see someone tweet..	non-moral	non-moral
8	Liberty and Justice for all? How about opp...	FairnessCheating	fairness
9	Took a long time, no? Doctors Strive to Do...	CareHarm	care
10	Yes RT @arthur_affect:Do ppl who change...	non-moral	non-moral
11	https://m.facebook.com/story.php?stor...	CareHarm	care

3.1.2 Moral Foundation Reddit Corpus

The Moral Foundation Reddit Corpus (MFRC) [15] should consist of 16,123 Reddit posts. The available dataset, downloadable from HuggingFace[1], starts off with 61,226 entries. Inspecting the downloadable dataset shows that every annotation is given its own entry, as shown in Table 3.3. 7399 of these annotations are duplicates, leaving 53,827 remaining entries. These entries also include pseudo-duplicates: the same annotator annotated the same post multiple times, but the labels are ordered differently, e.g. "Proportionality, Loyalty" and "Loyalty, Proportionality". There exist 292 pseudo-duplicates, resulting in 53,535 valid entries. To facilitate the process later on, all entries are condenses such that each remaining entry is a unique post, as shown in Table 3.4. This results in 17,886 unique entries.

Before assigning a single label for each entry, the fairness split [4] is reverted; the labels 'Equality' and 'Proportionality' are changed back into 'Fairness'. This step is taken to ensure that a better comparison between the Twitter and Reddit corpus can be made. For the sake of completion, the original labels are also stored in the dataframe.

Now, the most appropriate labels for each entry can be decided. The first step is majority voting, but this still leaves 3266 posts or 18.26% of the corpus with a tie for the most appropriate label. Trager et al. [15] also tasked the annotators to rate the confidence of their annotation. This confidence annotation can help with further distinguishing the most suited label. There are three levels of confidence available: Confident, Somewhat Confident, and Not Confident. Confident is used when a moral statement is clearly expressed and a single foundation can be attributed. In case of a not clearly expressed moral statement or sarcasm and when another foundation is vaguely present, Somewhat Confident is used. Lastly, when multiple foundations are equally present or the right foundation can not be attributed given the lack on context, Not Confident is attributed to the annotation. Using these interpretations for the levels of confidence, we assign the following strength factors for each level: factor 11 for Confident, factor of 5 for Somewhat Confident, factor of 2 for Not Confident, and a factor of 1 for a missing confidence level. These values for each confidence level are chosen such that a single Confident label will always be chosen over two Somewhat Confident labels. And the same goes for the relation between Somewhat and Not Confident. After confidence scaling, we still have ties for 2525 posts or 14.12% of the corpus, increasing the usable posts by 741. All remaining ties are handled automatically by Python's .count(max) algorithm, technically sorting all ties in alphabetical order. The resulting Reddit posts and their associated labels are structured like shown in Table 3.4 for easy access and processing later on, including a column 'moral-tie' indicating remaining ties for each label after confidence scaling.

As a result of the aforementioned processes, the resulting frequencies and label distributions differ from the ones reported by Trager et al. [15]. The obtained frequencies and distributions are reported in Table 3.5.

3.1.3 Annotator Agreement

Since attributing a moral label to posts is a highly subjective task, the rate of agreement between annotators is calculated to give a reliable insight in the universal agreement of the annotators. A typical way of representing the inter-annotator agreement is by using the Kappa statistic, which gives a general impression of the overarching agreement [23].

Both the Twitter Corpus and the Reddit Corpus have provided the results of the Fleiss' Kappa and PABAK (Prevalence-adjusted Bias-adjusted Kappa). Fleiss' Kappa allows calculating the rate of agreement between any number of annotators. PABAK also includes the distribution of unique labels to cover for the larger amount of possible variance between annotations. All Kappa scores are rated between 0 and 1, representing 0% and 100%

[1] https://huggingface.co/datasets/USC-MOLA-Lab/MFRC

Table 3.2: Frequency of labels in the Twitter corpus.

Foundation	N	Total [%]	Moral [%]
Care	2154	6.19	11.09
Fairness	2235	6.43	11.51
Authority	1365	3.93	7.03
Loyalty	2240	6.44	11.54
Purity	683	1.96	3.52
Harm	3521	10.12	18.13
Cheating	2999	8.62	15.44
Subversion	1755	5.05	9.04
Betrayal	1276	3.67	6.57
Degradation	1190	3.42	6.13
Moral	19,418	55.84	100
Non-moral	15,358	44.16	-
Total Labels	34,776	-	-

Table 3.3: Moral Foundation Reddit Corpus raw data.

text	subreddit	bucket	annotator	annotation	confidence
That particular par...	europe	French politics	annotator03	Non-Moral	Confident
That particular par...	europe	French politics	annotator01	Purity	Confident
That particular par...	europe	French politics	annotator02	Thin Morality	Confident
/r/france is pretty...	europe	French politics	annotator03	Non-Moral	Confident
/r/france is pretty...	europe	French politics	annotator00	Non-Moral	Somewhat Confident
/r/france is pretty...	europe	French politics	annotator02	Non-Moral	Confident
TBH Marion Le Pen...	neoliberal	French politics	annotator03	Non-Moral	Somewhat Confident
TBH Marion Le Pen...	neoliberal	French politics	annotator00	Thin Morality	Confident
TBH Marion Le Pen...	neoliberal	French politics	annotator02	Equality	Somewhat Confident
it really is a very un...	europe	French politics	annotator03	Non-Moral	Confident
it really is a very un...	europe	French politics	annotator04	Thin Morality	Confident
it really is a very un...	europe	French politics	annotator02	Non-Moral	Confident

Table 3.4: MFRC data stacked and labeled per post. 'text' contains the text of the reddit post. 'origin' list the indices of the original MFRC for each post. 'label' indicates the resulting label after majority voting and confidence scaling. 'moral-tie' lists whether the label was subject to a tie (1) or not (0).

	text	origin	label	moral-tie
0	That particular part of the debate is...	[0, 1, 2]	Non-Moral	1
1	/r/france is pretty lively, with it's own...	[3, 4, 5]	Non-Moral	0
2	TBH Marion Le Pen would be better. ...	[6, 7, 8]	Non-Moral	1
3	it really is a very unusual situation...	[9, 10, 11]	Non-Moral	0
4	The Le Pen brand of conservatism and...	[12, 13, 14]	Thin Morality	1
5	Macrons face just screams "I do not...	[15, 16, 17]	Non-Moral	0
6	Clinton lead polls by 4%, well within...	[18, 19, 20]	Non-Moral	0
7	Hey, fuck you. Us leftists will never...	[21, 22, 23]	Fairness	0
8	Clearly there were enough to affect...	[24, 25, 26]	Thin Morality	1
9	You are simplifying it. Islam is not the...	[27, 28, 29]	Care	0
10	Wow did not know all that! Maybe got...	[30, 31, 33]	Non-Moral	0
11	What planet are you on? Over 70% of...	[33, 34, 35]	Non-Moral	0

Table 3.5: Frequency of labels within the Reddit corpus after assigning labels.

Label	N	Total [%]	Moral [%]
Care	1843	10.30	24.88
Fairness	1987	11.11	26.82
Authority	778	4.35	10.50
Loyalty	434	2.43	5.86
Purity	319	1.78	4.31
Thin Morality	2047	11.44	27.63
Total Moral	7408	41.42	100.00
Non-Moral	10,478	58.58	-
Total labels	17.886	-	-

inter-annotator agreement respectively. Interpretations of the Kappa score may vary from source to source depending of the type of data that is being used [23]. McHugh proposes a rather strict interpretation, shown in Table 3.6, to ensure that a higher percentage of the data is seen as reliable [23]. The results of the calculated Kappa-scores are shown as heat maps in Figures 3.1 (MFTC) and 3.2 (MFRC), where a darker colour indicates a higher agreement and the lighter colour indicates lower agreement.

Observing the reported Fleiss' Kappa scores shows that for both the Twitter Corpus, Figure 3.1a, and the Reddit Corpus, Figure 3.2a, the inter-annotator agreement is generally minimal to weak. This can be explained by both the subjectivity of the task and the amount of available labels, 10 labels for the MFTC and 8 labels for the MFRC.

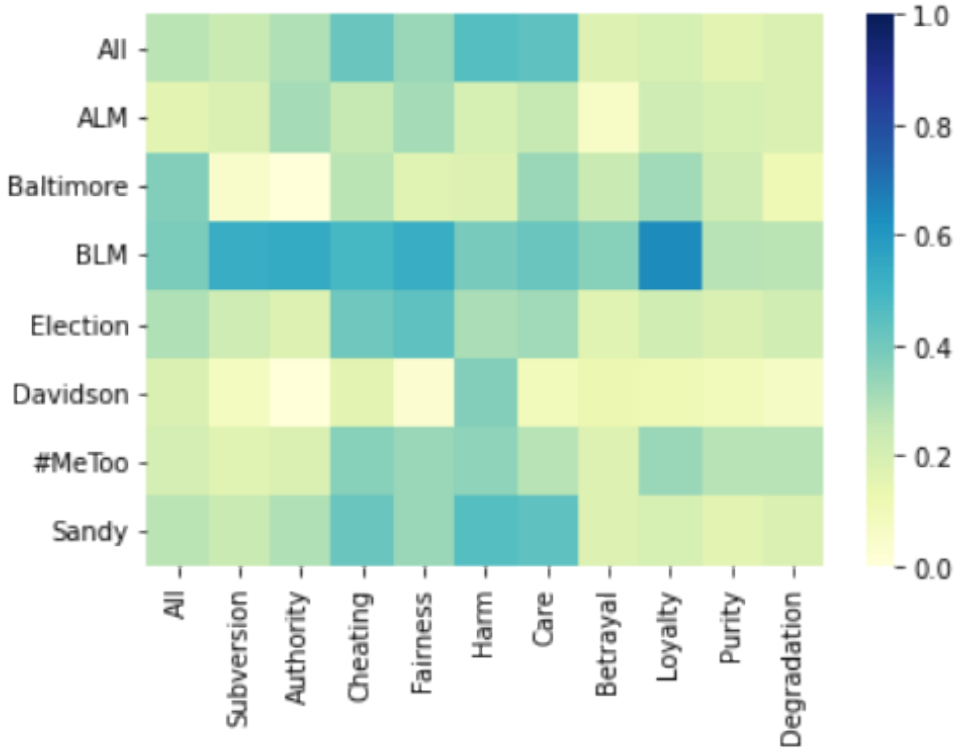
If the inter-annotator agreement is calculated using PABAK, and therefore adjusted to both prevalence and bias of labels, we see a higher degree of reported agreement. Where Fleiss' Kappa resulted in generally minimal, PABAK resulted in moderate to strong agreement. Big outliers are the Thin Morality and Non-Moral labels in the Reddit Corpus, showing only minimal agreement at best.

Table 3.6: Interpretation of Kappa-scores based on McHugh's [23] levels of agreement.

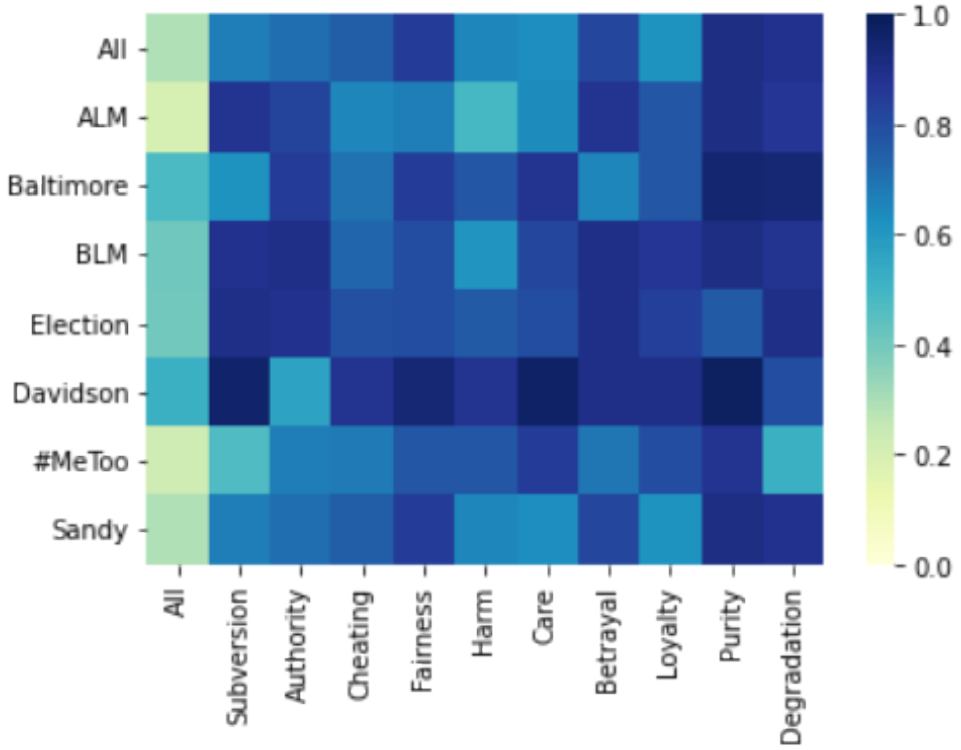
κ	Level of agreement	% Reliable data
-	-	-
.00 - .20	None	0 - 4%
.21 - .39	Minimal	4 - 15%
.40 - .59	Weak	15 - 35%
.60 - .79	Moderate	35 - 63%
.80 - .90	Strong	64 - 81%
> .90	Almost perfect	82 - 100%

3.2 Preprocessing

To prepare the data set for automatic extraction of moral foundations, the data needs to be cleaned of unnecessary information and simplified such that automatic processing steps are not faced with any problematic instances. This includes lowering all cases, removing stop words, removing punctuation, and removing URLs and HTML tags. For the Twitter corpus, this also includes removing twitter-handles and usernames. Python sees 'Word' and 'word' as different entries. Lowering all cases eliminates issues that can arise when capital

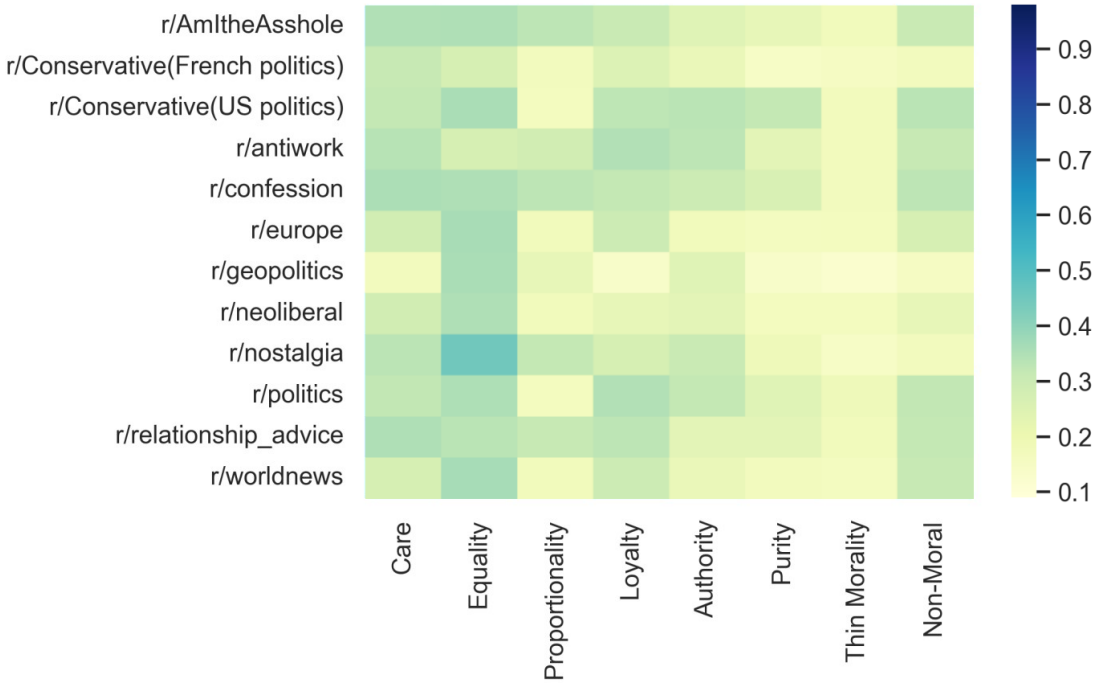


(a) Fleiss' Kappa

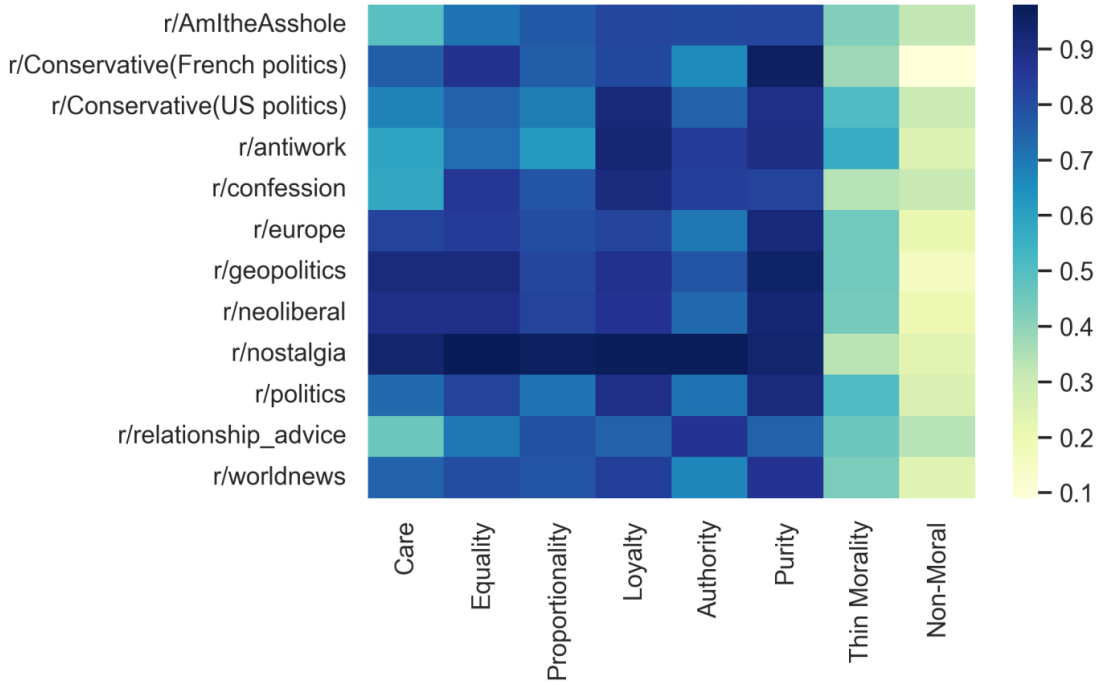


(b) PABAK

Figure 3.1: Annotator Agreement Heatmap for the Moral Foundation Twitter Corpus, created using data from Hoover et al. [16]. "All" merges all foundations on the x-axis and all topics on the y-axis.



(a) Fleiss' Kappa



(b) PABAK

Figure 3.2: Annotator Agreement Heatmap for the Moral Foundation Reddit Corpus, obtained from Trager et al. [15]. Sub-figure 3.2a has been adjusted to correctly reflect the annotator agreement on a similar scale as Sub-figure 3.2b.

letters are present. Stop words refer to words that occur often in any given text and therefore hold little to no information with regard to the semantics of a piece of text. Removing these stop words ensures that uniquely identifiable and relevant terms are retained. The NLTK.corpus package has a pre-made list of stop words and these include words like 'the', 'you', 'have', 'is', 'and', 'of', 'about' among many others. The final step is removing any punctuation, since this does not provide any information. The assigned labels cover the whole post and as such any distinction between individual sentences has disappeared. The NLTK Python library [24] provides the tools needed to easily apply these steps. The Tweet Tokenizer handles tokenization and lowering upper case letters at the same time. Table 3.7 shows the process of applying these steps to an example sentence, where each output is the input for the next step. Giving each element its own entry in the output lists facilitates the word embedding process.

Table 3.7: Result of preprocessing steps on an example sentence

Step	Output
Initial Sentence	"At eight o'clock on Thursday morning, Arthur didn't feel very good."
Tokenization	['at', 'eight', "o'clock", 'on', 'thursday', 'morning', ',', 'arthur', "didn't", 'feel', 'very', 'good', '.']
Stop word removal	['eight', "o'clock", 'thursday', 'morning', ',', 'arthur', 'feel', 'good', '.']
Punctuation removal	['eight', "o'clock", 'thursday', 'morning', 'arthur', 'feel', 'good']

Another commonly used pre-processing technique is lemmatization, which is converting a word into its smallest core element. For example, 'liking' and 'liked' are converted to 'like' and 'wrote' and 'writing' are converted into 'write'. Context is an important factor for lemmatization, since in some cases simply taking the lemma is not the right approach. 'writing' could also refer to that which is written down instead of the active verb. Since context is an important factor for lemmatization, we did not include it because it is expected that the Twitter corpus is subjected to contain less context to accommodate the character limit per post.

3.3 Text Embedding

After the aforementioned preprocessing steps, every post will be represented as a vector using the word representations from Gensim's pre-trained Word2Vec (W2V) models [8, 25] and the extended Moral Foundation Dictonary [10]. The selected Gensim model is 'glove-twitter-200', because this model is trained on Twitter instead of Wikipedia or Google News. Both the Moral Foundation Twitter and Reddit Corpora are derived social media platforms and will likely use different language and relations from encyclopedia or news outlet style like Wikipedia or Google News, respectively. Additionally, it has the highest amount of dimensions out of the pre-trained Twitter models. The 200 indicates the length and in turn the amount of dimensions for each vector. A greater amount of dimension allows for more elaborate relations between vectors, resulting in better performances [7].

For each post, the resulting vector will be generated by summing the word vectors for every word, skipping the word if no associated word vector is available. This sum is then divided by the number of words in the processed post, resulting in a normalised vector.

This process is also described by this notation:

$$V_{post} = \frac{\sum_{i=1}^{n} v_i}{n} \begin{cases} n & = \text{number of words remaining after processing} \\ v_i & = \text{associated word vector} \end{cases}$$

Using the example sentence from Table 3.7 results in:

$$V_{example} = \frac{\sum(v_{eight}, v_{o'clock}, v_{thursday}, v_{morning}, v_{arthur}, v_{feel}, v_{good})}{7}$$

When pre-processing steps result in a post not containing any words, then the resulting vector becomes a zero vector with the length to match the text representations; a length of 200 for the Gensim W2V model and a length of 5 for the eMFD.

After applying embedding techniques, some entries might not contain any information, because none of the remaining words are present in the word embedding vocabulary. This is unlikely for a big model like Gensim's Glove-twitter-200, but will happen more often for smaller dictionaries like the eMFD. These empty entries are removed to avoid skewing the training process.

4 METHOD

This chapter explains the experiments that will generate the information required to answer the research questions, using the data that was prepared in chapter 3 and following the schematic in Figure 4.1. The measurement metric for the experiments will be the F-score to reflect a balance between accuracy and precision.

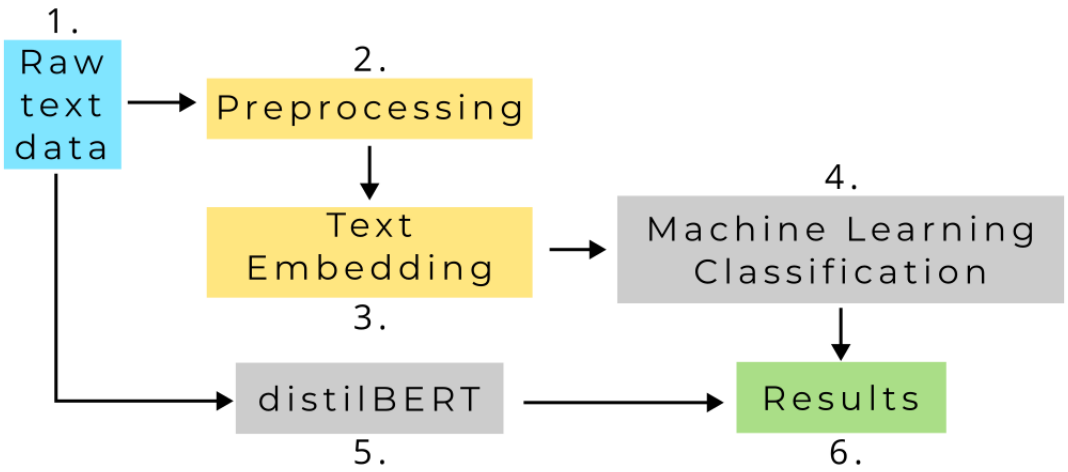


Figure 4.1: Schematic overview of the taken to generate the results.

4.1 Experiment description

The non-moral label is the dominating label within both data sets, as evident in Tables 3.2 and 3.5. To facilitate the process of moral classification, all experiments are run in two variations: 1) binary classification; classifying all moral labels versus the non-moral label and 2) moral-only classification; excluding the non-moral label and performing the classification within the remaining moral labels. This is to ensure that lower frequency labels, such as Purity in both MFTC, Table 3.2, and MFRC, Table 3.5, are less likely to be misclassified.

For moral-only classification, all posts labeled 'non-moral' will be removed, keeping the original moral labels intact. All other steps are exactly the same as for binary classification, as scikit-learn's Support Vector Machine (SVM) and Logistic Regression models automatically handle multi-class classification.

4.1.1 Text Representation

To test the difference in effectiveness for the selected text representations, as described in Section 3.3, the same steps will be executed twice: first using Gensim's W2V representation, followed by the eMFD representation. Now, another split will be made into binary

classification and moral-only classification.

Starting with binary classification first, all posts that are labeled with one of the moral traits will simply be relabeled as 'moral'. Next, all posts that have a resulting 0 vector after applying embedding techniques are removed, since they contain no information to train on. Selected classification models are scikit-learn's SVM and Logistic Regression.

4.2 Traditional versus Modern techniques

The previous experiment uses traditional classification techniques. To get a more complete overview, a classification using modern techniques is performed. BERT [19] is a high-end classifier with built-in embedding and has numerous pre-trained variants and models. DistilBERT[1] [26], a smaller variant of BERT, is selected for its speed and performance, reaching 95% of BERT's language understanding and being 60% faster [26]. Because of these built-in features and tools, the input for distilBERT is the raw text data, like shown in the 'text'-columns in Tables 3.1 and 3.4. To align with the previous experiment, both a binary classification and a moral-only classification will be performed.

The model is fine-tuned for 10 epochs, using a learning rate of 2e-5, and a batch size of 8. These parameters are selected based on similar parameters for experiments run by Bulla et al. [18] and Trager er al. [15].

4.3 Cross data set classification

Cross data set classification tries to extrapolate from given information, making it a desired tool in cases where (labeled) data is limited. With the MFTC and MFRC, we can easily verify if either are a suitable data set to extrapolate from.

For this experiment, the binary classification works similar to the previous experiment; non-moral is kept as is and all other labels will be relabeled as moral. Properly performing morals-only classification requires a bit more work, since the two datasets have different pools of labels. The MFTC has labels for both virtues and the vices, whereas the MFRC makes no difference between the two. Therefore, relabeling the vices from the MFTC into their virtues should solve this issues. For example, both the label 'care' and 'harm' are now labeled as 'care'. The 'Thin Morality' morality label in the MFRC has no similar counterpart within the MFTC. Therefore, all 'Thin Morality' labels have been removed from the MFRC for morals-only classification. However, 'Thin Morality' is still a valid moral label and will still be included during binary classification. This realignment of labels results in new frequencies for all relevant labels, shown in Table 4.1.

The best performing text representation and classifier from the previous experiment will be used for this experiment as well.

To perform this experiment, one model will be trained on the Twitter corpus and tested on the Reddit corpus. Then, another model will be trained on the Reddit corpus and tested on the Twitter corpus. Because the training and testing sets are independent, no stratified K-fold is required for this experiment.

4.4 Experimental Setup

4.4.1 Classification Distribution

The data set is trained and tested with a split based on stratified K-Fold, with $K = 5$. K-Fold is a cross-validation technique where each resulting subset, or fold, is used once for testing and $K - 1$ times for training. The stratified variation retains the percentage for each label within each fold, which is ideal for unbalanced data sets.

[1] https://huggingface.co/docs/transformers/model_doc/distilbert

Table 4.1: Resulting label frequencies after matching the MFTC and MFRC labels.

	MFTC			MFRC		
Label	N	Total [%]	Moral [%]	N	Total [%]	Moral[%]
Care	5675	16.32	29.23	1843	10.30	34.38
Fairness	5234	15.05	26.95	1987	11.11	37.06
Authority	3120	8.97	16.07	778	4.35	14.51
Loyalty	3516	10.11	18.11	434	2.43	8.10
Purity	1873	5.39	9.65	319	1.78	5.95
Moral	19,418	55.84	100.00	*7408	*41.42	100.00
Non-moral	15,358	44.16	-	10,478	58.58	-

* Thin Morality is kept within the Moral labels for binary classification, n = 2047.

4.4.2 Evaluation of models

To score the success rate of the algorithms, the F-score will be used, representing the harmonic mean of precision and recall. As such, both precision and recall contribute equally and are both represented by using the F-score. The formula for the F-score is denoted as:

$$F-score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad \text{or} \quad F-score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

$$\text{with:} \quad precision = \frac{TP}{TP + FN} \ \& \ recall = \frac{TP}{TP + FP}$$

$$\text{and:} \quad TP = TruePositive, FP = FalsePositive, FN = FalseNegative$$

The F-score can range from 0 to 1; where 1 indicates both perfect precision and recall and 0 indicates either precision or recall being equal to zero.
The resulting F-score will be compared to random classification, where the baseline reflects the percentage of occurrence of the label within its respective data set. These scores are already portrayed in Tables 3.2, 3.5, and 4.1. Where possible and relevant, the resulting F-score will also be compared results from similar experiments performed in literature, mainly results from the MFTC [16], MFRC [15] and Bulla et al [18].

4.4.3 Implementation details

All scripts are programmed in Python, because it is widely available and free to use. Python also has widely available materials that support machine learning and deep learning, such as scikit-learn [27], TensorFlow [28], PyTorch [29], and Keras [30]. Entire pre-trained models, such as (distil)BERT [19, 26], are also generally available. The Natural Language ToolKit (NLTK, [24]) helps with handling human language data, just like the data available in the corpora, providing tokenizers and lists of stopwords. The Pandas library [31] helps with generating easy to use spreadsheets and saving them as csv-files.

Most scripts are run locally using Jupyter Notebooks and a Intel Core i7-10750H CPU with 16GB RAM. Scripts using distilBERT is run in Google Colab [2] using its NVIDIA T4 Tensor Core GPU[3].

[2] `https://colab.research.google.com/`
[3] Datasheet: `https://www.nvidia.com/en-us/data-center/tesla-t4/`

5 RESULTS

The results of the experiments will be shown here. The baseline measure will be random classification comparing the resulting F-scores to the percentage of the label within a dataset. All experiments described in Chapter 4 will be answered in order, to reflect the sub-research questions; different text representations, machine learning compared to deep learning, and suitability for cross data set prediction.

5.1 Text representations

The first research question deals with investigating the effectiveness of text representations. This research uses one of GloVe's [7] pre-trained Word2Vec [8] representations, a general text model, and the extended Moral Foundation Dictionary (eMFD, [10]), a domain specific model. As such, we start with investigating these differences, beginning with the results of binary classification in Table 5.1. Looking at the W2V representation, all baselines are beaten with sufficient margin. Looking at the Twitter corpus, we see slightly better results for classifying the moral tweets compared to the non-moral tweets. For Reddit this is the other way around, where non-moral posts are classified with greater success. The eMFD representation reflects these results with a severe drop in success for non-moral in Twitter and moral in Reddit.

This could be a result of the imbalance of the label distributions. However, this does not explain the severe drop in f-score when using eMFD. Instead, it could also indicate that Twitter is more prone to moral language and Reddit more prone to non-moral language.

Continuing with the moral-only classification, we take a look at Tables 5.2, 5.3, and 5.4, covering the results for Twitter's virtues, Twitter's vices, and the Reddit corpus, respectively. Beginning with Twitter's virtues, Table 5.2, it immediately becomes clear that eMFD is bad at classifying virtues in Twitter. It has more success at classifying vices, Table 5.3, beating the baselines if some successful classification is present. Even then, the eMFD is greatly outclassed by W2V, which beats all baselines with greater margins. For Reddit, Table 5.4, we see similar results. The eMFD is unable to classify most foundations correctly, and is outclassed in cases where it performs some successful classification.

Inspecting the W2V representation results, Twitter's baselines are beaten for both the virtues and the vices. However, the resulting score for classifying ”Care”, .160-.210, is very mediocre. In comparison, the score for ”Purity”, approximately .310, is not great either, but can still be explained by the relatively small percentage in the data set. At the same time, ”Degradation” also performs slightly worse compared to the other vices, approximately .415 as opposed to > .500, hinting at some difficulties to classify the foundation of ”Purity”. Further inspecting the W2V representations of the Reddit corpus results, Table 5.4, we see that both ”Care” and ”Fairness” do not beat the baseline and ”Authority” only beats it with minor success. The failure of classifying ”Care” in both Twitter and Reddit can indicate that this foundation is prone to using non-moral language.

Table 5.1: Resulting F-scores for binary classification including baseline threshold and standard deviation. Results from Bulla et al. [18] and Trager et al. [15] are included for comparison. Best results per column are marked in bold.

		Twitter		**Reddit**	
		Moral	Non-moral	Moral	Non-moral
	Baseline	.558	.442	.414	.586
SVM	W2V	.785 ± .062	.714 ± .096	.614 ± .082	.769 ± .027
	eMFD	.729 ± .011	.129 ± .051	.142 ± .055	.732 ± .008
LogReg	W2V	.782 ± .064	.714 ± .095	.619 ± .079	.769 ± .028
	eMFD	.729 ± .010	.126 ± .054	.139 ± .054	.733 ± .008
distilBERT		**.888**	**.847**	.703	**.820**
Literature		.85 [18]	-	**.73** [15]	-

5.2 Traditional versus modern techniques

The second research question deals with the implications of traditional methods and modern methods. From sub-research question 1, Section 5.1, it became evident that these results based on the eMFD are not worthwhile for further investigation. As such, we will focus on the W2V representations and the differences between SVM, Logistic Regression and (distil)BERT. Once again starting with binary classification from Table 5.1, we see that SVM and Logistic Regression have very similar results; significantly beating the baseline and performing with acceptable success. distilBERT, on the other hand, has roughly a .100 higher f-score, highlighting greater success rates and thus effectiveness.

Inspecting the differences for SVM and LogReg for Twitter, Tables 5.2 and 5.3, no big differences are observed. Logistic Regression seems to ever so slightly take the lead when it comes to foundations with lower success rates (approximately < .500), such as ”Care” and ”Degradation”. The same statements can be made for Reddit, Table 5.4. Within these two traditional methods, there is hardly a significant difference in results. Comparing these results to distilBERT’s output, we see that this approach ranks better across the board. Starting with Twitter, Tables 5.2 and 5.3, distilBERT provides more consistent results for all virtues, being only outclassed by the SVM method for the Fairness and Subversion dimensions. Out of the remaining dimensions, only Loyalty, Harm, and Degradation start to come close (within .10 range) to the results of distilBERT. For the Reddit corpus, Table 5.4, distilBERT’s results prove less reliable, outclassing only the foundations of Care, Fairness, and Authority. Using distilBERT to classify Purity provides very similar results to the SVM and Logistic Regression approach, but is outperformed on Loyalty and Thin Morality.

5.3 Cross Data Set Classification

Tables 5.5 and 5.6 show the results for cross data set classification. Starting with binary classification, Table 5.5, it is evident that the baseline is beaten. Comparing the cross data set results to the within data set results, similar f-scores (within .050 margin) are achieved. It is notable that the smaller partition of the two labels is classified slightly less successfully, similar to the within data classification. This is unexpected since the training set has the other label being the larger of the two compared to the testing set. This likely indicates that the Twitter data set is more prone to using moral language whereas Reddit posts likely

Table 5.2: Resulting F-scores for morals-only classification of the virtues from the Moral Foundation **Twitter** Corpus, including baseline threshold and standard deviation. Best results per foundation are marked in bold.

Virtues		Care	Fairness	Authority	Loyalty	Purity
	Baseline	.111	.115	.070	.115	.035
SVM	W2V	.163 ± .101	.682 ± .073	.522 ± .140	.605 ± .083	.310 ± .074
	eMFD	0 ± 0	.381 ± .083	0 ± 0	0 ± 0	0 ± 0
LogReg	W2V	.214 ± .077	.664 ± .069	.515 ± .123	.601 ± .061	.316 ± .067
	eMFD	0 ± 0	.373 ± .076	0 ± 0	0 ± 0	0 ± 0
	distilBERT	.651	.519	**.719**	**.664**	**.674**
	Bulla et al. [18]	**.75**	**.77**	.60	.66	.57

Table 5.3: Resulting F-scores for morals-only classification of the vices from the Moral Foundation **Twitter** Corpus, including baseline threshold and standard deviation. Best results per foundation are marked in bold.

Vices		Harm	Cheating	Subversion	Betrayal	Degradation
	Baseline	.181	.154	.090	.066	.061
SVM	W2V	.538 ± .158	.544 ± .071	**.583 ± .045**	.575 ± .050	.410 ± .128
	eMFD	.392 ± .105	.227 ± .066	.265 ± .045	.453 ± .038	0 ± 0
LogReg	W2V	.533 ± .152	.531 ± .073	.569 ± .028	.580 ± .044	.425 ± .128
	eMFD	.406 ± .130	.266 ± .068	.265 ± .039	.464 ± .040	0 ± 0
	distilBERT	.635	**.697**	.409	**.742**	.451
	Bulla et al. [18]	**.68**	.67	.47	.55	**.50**

Table 5.4: Resulting F-scores for morals-only classification of the Moral Foundation **Reddit** Corpus, comparing Gensim's Word2Vec and the eMFD text representation. Best results per foundation are marked in bold.
*****Estimation based on the f-scores for Equality and Proportionality and their distribution within the dataset.

	Care	Fairness	Authority	Loyalty	Purity	Thin Morality
Baseline	.249	.268	.105	.059	.043	.276
SVM W2V	.113 ± .069	.211 ± .073	.193 ± .044	**.527 ± .089**	.460 ± .042	**.563 ± .026**
eMFD	0 ± 0	0 ± 0	0 ± 0	.011 ± .009	.388 ± .010	.348 ± .019
LogReg W2V	.142 ± .080	.255 ± .082	.222 ± .056	.520 ± .104	**.468 ± .039**	.558 ± .026
eMFD	0 ± 0	0 ± 0	0 ± 0	0 ± 0	.393 ± .010	.348 ± .025
distilBERT	**.59**	**.60**	.26	.46	**.50**	.38
MFRC [15]	**.59 ± .02**	.46 ± .02*****	**.35 ± .05**	.43 ± .04	.48 ± .07	.34 ± .04

contain more neutral or non-moral language.

Stepping over to the moral-only cross classification, portrayed in Table 5.6, all baselines are beaten. This signifies some degree of usefulness for cross data classification within the moral domain.

The first notable result is the successful classification of 'Care', for both Twitter and Reddit data sets, beating the within data classification significantly, where this foundation scored relatively low and failed to beat the Reddit baseline. This is also the case for 'Fairness' and 'Authority' when classifying for Reddit. All other foundations score worse during cross data classification than during within data classification, which is expected.

These results indicate that cross data classification is only somewhat possible, but results will be inferior to within data classification being more than .100 lower. Classifying the "Care" foundation is successful across data sets, with an f-score of approximately .600 for both corpora. In addition, the "Fairness" foundation is also still classified to an acceptable degree, f-score of .500, when testing on the Reddit Corpus.

Table 5.5: Resulting F-score for binary cross data set classification, using W2V representation and SVM classifier. Best results per column are marked in bold.

Test set	**Twitter**		**Reddit**	
Train set	Reddit		Twitter	
	Moral	Non-moral	Moral	Non-moral
Baseline	.558	.442	.414	.586
Within data	**.785 ± .062**	**.714 ± .096**	**.614 ± .082**	**.769 ± .027**
Results	.750	.662	.516	.734

Table 5.6: Resulting F-score for moral-only cross data set classification, using W2V representation and SVM classifier. Results are shown for the data set that served as the testing data. The new baselines were previewed in Table 4.1. Best results per foundation are marked in bold.
*****Twitter's within data has separately been generated to function as a comparison for this experiment.

Evaluation set		Care	Fairness	Authority	Loyalty	Purity
Twitter	Baseline	.292	.270	.161	.181	.097
	Within data*	.490±.125	**.657±.045**	**.504±.056**	**.650±.071**	**.420±.180**
	Results	**.603**	.412	.383	.379	.282
	BERT [15]	.53	.35	.38	.38	.28
Reddit	Baseline	.343	.371	.145	.081	.060
	Within data	.113±.069	.211±.073	.193±.044	**.527±.089**	**.460±.042**
	Results	**.574**	**.507**	**.332**	.193	.254
	BERT [15]	.43	.34	.31	.32	.34

6 DISCUSSION

During this research, we encountered issues and made decisions based on the materials at hand. This chapter serves to elaborate on the results and the implications of these decisions. This naturally leads to both limitations and recommendations for future work, which both will be addressed.

6.1 Research Questions

6.1.1 To what extent do different text representations affect the results?

Initially, it was expected that a specialised dictionary, like the extended Moral Foundation Dictionary (eMFD), would prove to be an effective embedding technique. However, the results showed that the eMFD was greatly outclassed by Gensim's pre-trained Word2Vec model. The eMFD can still reliably classify moral against non-moral posts, provided that the right target, moral or non-moral, is chosen. For moral-only classification, only vices saw some successful classifications. The results from using Gensim's W2V [8] as word embedding shows that a large general model surprisingly outperforms a specialised, but much smaller, model.

6.1.2 What are the implications of using deep learning or machine learning techniques and how do they compare in performance and usability?

Comparing the results of machine learning techniques, Support Vector Machines and Logistic Regression, against a deep learning classifier, distilBERT, show that deep learning classification generally provides better and especially more consistent classification for all moral foundations. However, it is worth mentioning that machine learning techniques still provide acceptable results while being significantly faster and less demanding in terms of processing power. Improving on pre-processing or text embedding can help reducing the limitations and make machine learning techniques a more competitive alternative to deep learning.

In general, we find that binary classification - moral against non-moral - shows reliable results regardless of text embedding or classifier. As such, it can be applied without much drawbacks to get an initial insight or separate non-moral posts from a data set.

One final remark is the very big difference in demand for processing power. All machine learning tasks are executed on a local CPU[1] and take only a matter of seconds to complete or up to a couple of minutes to generate the different text representations, the latter of which can be saved and easily accessed again for other applications. The speed and relatively low requirements allow for elaborate testing and experimentation without worrying too much about optimizing code or extensive downtime between experiments. To run the experiments using (distil)BERT, a powerful GPU is desired. During this research, Google Colab's T4 GPU has been used when running distilBERT. Whilst very powerful and well-suited for the task, its access and availability is limited to about 3 hours per 24 hours

[1] CPU specs: Intel Core i7-10750H, 16GB RAM

with upgrades and extensions available behind a paywall. Even with such a powerful tool, a full classification task took roughly 15 minutes per dataset per epoch.

6.1.3 How well do the available corpora lend themselves for cross data set training and testing?

Binary cross data set classification performs only slightly worse compared to within corpus classification, still achieving very fair scores. This indicates that both corpora are well suited for cross data set classification, provided that the right target class, which is either moral or non-moral, can be determined beforehand. Morals-only cross data set classification is more limited in its usability, showing only promising results for the foundation of ”Care”. Lastly, machine learning techniques resulted in very similar scores compared to BERT based on literature [15], once again highlighting the potential for machine learning techniques for classifying moral foundations.

6.1.4 To what extent can Natural Language Processing techniques identify Moral Foundations in text?

Combining all findings for the sub research questions, we find that detecting moral foundations from text is a reasonably challenging task. Binary classification yields good results (f-score > .700) for both the Twitter and the Reddit corpora and cross data set classification.

For moral-only classification, we start to see more limited results. The Moral Foundation Twitter Corpus shows generally promising results across the board. Using distilBERT provides better results, but in most cases the machine learning approaches are within a .150 margin. Only the dimensions of ”Care” and ”Purity” show significant difference between machine and deep learning methods, whereas ”Degradation” is difficult to predict regardless of method.

The Moral Foundation Reddit Corpus shows no clear distinction of successful classification. While, distilBERT is vastly superior for detecting the ”Care” and ”Fairness” foundations, SVM and Logistic Regression are better in detecting ”Loyalty” and ”Thin Morality”. Neither the machine learning nor the deep learning approach can successfully classify ”Authority”. One possible explanation is that the MFRC combines both the virtues and the vices into a single foundations, possibly losing some nuance, whereas the MFTC keeps the distinction between virtues and vices.

The materials at hand are well suited as a tool to gain general insight of moral presence a piece of text. However, the materials are still not good enough to make reliable classifications on a functional level, as the results for cross data set classification only achieve an f-score up to .600 in the best case.

6.2 Limitations

The biggest limitation of this research is that the classification results are only as good as the data on which a classification model has been trained. Both the MFTC and MFRC are manually annotated and, in most cases, have a moderate to strong level of inter-annotator agreement. This also means that a resulting level of reliability for both the annotation and the classification tasks is constrained by the quality of the annotations. The Moral Foundation Theory wasn't created to make strong distinctions between different moral statements, but to help get an insight in the underlying intuitions of a conversation partner.

6.2.1 Preprocessing

Currently, lemmatization has not been applied in the preprocessing steps. This is expected to not be too much of an issue, since the word representation methods, especially Gensim's

Word2Vec [8], already have a very elaborate database making it unlikely that keywords are missed. The eMFD contains a reasonable amount of word variations, where each variation has different moral scores. For example the word ”vote” and its variations, as shown in Figure 6.1, score differently across the foundations. While these words appear very closely related, the resulting scores can vary greatly. Applying lemmatization would nullify the differences that have been assigned to each variation of a word such as ”vote”. In addition, lemmatizers can be heavily dependent on context, which is likely to be missing when dealing with tweets.

word	care_sent	fairness_sent	loyalty_sent	authority_sent	sanctity_sent
vote	0.056665	-0.065617	0.033044	0.056184	0.031260
voted	-0.088461	-0.042170	0.006534	0.034449	-0.120554
voter	-0.307019	-0.456126	-0.349762	-0.438567	-0.347300
voters	-0.112522	-0.055480	0.041426	-0.037075	-0.057200
votes	-0.249778	-0.130302	-0.077500	-0.017110	-0.064947
voting	-0.113797	-0.161890	-0.283620	-0.203786	-0.209235

Figure 6.1: All variations of the word ”vote” within the extended Moral Foundation Dictionary [10]

6.2.2 New foundations

Newer iterations elaborate on the original Moral Foundation Theory by expanding the existing foundations. Notable are the addition of ”Liberty” [5] and Fairness split into ”Equality” and ”Proportionality” [4]. Both these examples have been ignored or reverted in case of the Reddit corpus [15] during this research. This has been done since other supporting materials do not cover these improvements. Especially, because similar supporting materials are not yet available. These additions are likely very fitting for the MFT as a whole, but classification and prediction of these morals is still limited by the amount of annotated data. At the time of writing, only Trager et al. [15] have applied updates of the MFT in creating a labelled dataset.

When more annotated data sets based on the MFT are created, inclusion of these new additions could lead to different results and conclusions from the ones obtained in this research and other cited works.

6.3 Future Work

6.3.1 Improved Text Embedding

In this research, we applied average text embedding, where all words are seen independent from each other and any potential context is ignored. To improve on this, methods such as verb embedding [14] or context embedding, for example sentenceBERT [32], can be implemented. Verb embedding deals with specifying the object belonging to certain verbs. Context embedding tries to form links in order to retain the relations between words in a sentence.

6.3.2 Thematic alignment

Both the MFTC [16] and the MFRC [15] also included a label for the theme or page of retrieval. For example #AllLivesMatter for Twitter or the ”europe” subreddit of Reddit. Observing results from their respective papers, it is evident that a subset generally performs better than the whole corpus at once. Taking this into account, it can be valuable to train a classification model based on thematic alignment. For example, training on the Hurricane Sandy subset of the MFTC could prove beneficial when predicting or classifying moral sentiment on social media during hurricanes or other natural disasters.

6.3.3 Ranked Classification

Because the Moral Foundation Theory was created to help get an insight in the moral standpoint of a conversation partner, a logical next step would be to expand on the prediction of the classification. This expansion could be focused on returning a ranked list of the top foundations. This immediately covers statements that are morally ambiguous or statements where multiple foundations are present to similar degrees.

7 CONCLUSION

This thesis aimed to investigate to what extent Natural Language Processing techniques can identify Moral Foundations in text. We compared different embedding and classification techniques and ended the research with investigating cross data set classification.

We found that the dedicated extended Moral Foundation Dictionary is not well suited for classification of moral foundations. A general pre-trained model, like Gensim's Word2Vec or BERT's built-in representation model, are much more reliable.

While the results from BERT outclass the simpler machine learning methods, it does come with higher requirements for both operating hardware and time required. For strict classification and result optimisation, going the extra mile to meet the requirements for classifying using a deep learning model like BERT is likely worthwhile. Whereas if the goal were somewhat simpler, for example getting insights for personal use, using machine learning methods is more than suitable for the task at hand.

Binary cross data set classification performs only slightly worse compared to within corpus classification, still achieving very fair scores. This indicates that both corpora are well suited for cross data set classification, provided that the right target class, which is either moral or non-moral, can be determined beforehand. Morals-only cross data set classification is more limited in its performance.

The Natural Language Processing tools and approached used during this research successfully provide a general insight in detecting moral foundations in text, but lack the capability for reliable detailed classification.

7.1 Contributions

The main contribution is the implications of cross data set classification. With limited data sets available, cross data set classification helps overcoming these limitation by branching out into domains without annotated data. Even though the results for moral cross data classification are generally mediocre, it still sufficiently outclasses random classification. Especially the foundation of "Care" is well suited for cross data set classification. The results for binary cross data classification are more than sufficient to warrant cross data classification in other domains. This facilitates the initial steps for branching into other content or even helping with generating and annotating more corpora for classification of moral foundations.

Secondly, this research gives the insight that applying machine learning techniques are not inherently an invalid option for classification of moral foundations in text. They are outclassed by a deep learning classifier, but they still show comparable results at a fraction of the time and processing power.

7.2 Acknowledgements

I'd like to thank dr. Lorenzo Gatti for hydrating the Moral Foundation Twitter Corpus and simplifying the labelling process. The MFTC [16] only included reference codes to the posts and not the posts themselves.

REFERENCES

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

[2] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, vol. 20, no. 1, p. 98–116, 2007.

[3] J. Graham and J. Haidt, "Other materials | moral foundations theory." `https://moralfoundations.org/other-materials/`. Retrieved 10 May 2022.

[4] M. Atari, J. Haidt, J. Graham, S. Koleva, S. Stevens, and M. Dehghani, "Morality beyond the weird: How the nomological network of morality varies across cultures," *Journal of personality and social psychology*, vol. 125, pp. 1157–1188, 08 2023.

[5] R. Iyer, S. Koleva, J. Graham, P. Ditto, and J. Haidt, "Understanding libertarian morality: The psychological dispositions of self-identified libertarians," *PLoS ONE*, vol. 7, no. 8, 2012.

[6] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, 12 2009.

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1532–1543, 2014.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[9] J. A. Frimer, R. Boghrati, J. Haidt, and M. Dehghani, "Moral foundations dictionary for linguistic analyses 2.0." Unpublished Manuscript, 2019.

[10] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," *Behavior Research Methods*, vol. 53, no. 1, p. 232–246, 2020.

[11] 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. 1029–1046, 2009.

[12] O. Araque, L. Gatti, and K. Kalimeri, "MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction," *Knowledge-Based Systems*, vol. 191, 3 2020.

[13] M. C. Pavan, V. G. dos Santos, A. G. J. Lan, J. Martins, W. R. dos Santos, C. Deutsch, P. B. Costa, F. C. Hsieh, and I. Paraboni, "Morality classification in natural language text," *IEEE Transactions on Affective Computing*, 10 2020.

[14] J. Y. Xie, G. Hirst, and Y. Xu, "Contextualized moral inference," *Computing Research Repository*, 8 2020.

[15] J. Trager, A. S. Ziabari, A. M. Davani, P. Golazizian, F. Karimi-Malekabadi, A. Omrani, Z. Li, B. Kennedy, N. K. Reimer, M. Reyes, K. Cheng, M. Wei, C. Merrifield, A. Khosravi, E. Alvarez, and M. Dehghani, "The moral foundations reddit corpus," *Computing Research Repository*, 8 2022.

[16] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, G. Moreno, C. Park, T. E. Chang, J. Chin, C. Leong, J. Y. Leung, A. Mirinjian, and M. Dehghani, "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[17] B. Kennedy, M. Atari, A. M. Davani, J. Hoover, A. Omrani, J. Graham, and M. Dehghani, "Moral concerns are differentially observable in language," *Cognition*, vol. 212, p. 104696, 2021.

[18] L. Bulla, S. D. Giorgis, A. Gangemi, L. Marinucci, and M. Mongiovì, "Detection of morality in tweets based on the moral foundation theory," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13810 LNCS, pp. 1–13, 2023.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, June 2019.

[20] M. Wilson, "Why is Twitter called X now? Elon Musk's rebrand explained and where it's going next — techradar.com." https://www.techradar.com/computing/social-media/why-is-twitter-now-x-elon-musks-rebrand-explained-and-where-x-is-going-next, 2023. [Accessed 21-01-2024].

[21] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, no. 1, p. 512–515, 2017.

[22] S. Perez, "Twitter's doubling of character count from 140 to 280 had little impact on length of tweets.." Techchrunch, https://tcrn.ch/2zi0r02. Retrieved 2 May 2022.

[23] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, p. 276, 2012.

[24] "NLTK :: Natural Language Toolkit." https://www.nltk.org/. Accessed: 2024-06-03.

[25] "models.word2vec – Word2vec embeddings – gensim." https://radimrehurek.com/gensim/models/word2vec.html. Accessed: 2024-06-03.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiV*, vol. abs/1910.01108, 2019.

[27] “scikit-lean: machine learning in Python.” https://scikit-learn.org/stable/index.html. Accessed: 2024-06-03.

[28] “TensorFlow.” https://www.tensorflow.org/. Accessed: 2024-06-03.

[29] “PyTorch.” https://pytorch.org/. Accessed: 2024-06-03.

[30] “Keras: Deep Learning for Humans.” https://keras.io/. Accessed: 2024-06-03.

[31] “pandas - Python Data Analysis Library.” https://pandas.pydata.org/. Accessed: 2024-06-03.

[32] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Association for Computational Linguistics, Nov. 2019.