Values from texts

Francesco Palmisano 1004694 francesco.palmisano2@studio.unibo.it
Primiano Arminio Cristino 1004689 primiano.cristino@studio.unibo.it
Università degli studi di Bologna

***Abstract*—The construction of schemes for the annotation of corpora is a task in the field of data classification, concerning theoretical assumptions of the concept to be annotated. It defines what kind of data information must be annotated. However, explaining the annotations is challenging, in particular for AI systems. This paper will discuss the semantic relation between Haidt values [1] and texts and it tries to explain the reason of this relation. The purpose of this paper is twofold: (i) verify possible differences in the value detection by confronting the original Social Chemistry dataset with the automatic frame-based extraction; (ii) investigate the distribution of lexical value triggers respect to Glove model.**

I. Introduction

This paper [2] aims to explain the moral concern's polarity based on the Haidt value by analysing patterns in the Resource Description Framework (RDF) [3] of text corpora, contained in the Social Chemistry dataset [4]. In section 2 the paper describes the preprocessing steps which has been developed on the Social Chemistry dataset. In Section 3, there will be explained the analysis on the RDF of text corpora and the extraction of Haidt values. The frame based extraction in fact, being explainable in its keeping track of the triggering localisation, allows a more precise analysis than considering the original text chunk entry from Social Chemistry.

II. preprocessing

Jonathan Haidt proposed that human beings have six different areas of moral concern: care, fairness, liberty, loyalty, authority, and purity, which they call "Foundations."
The Social Chemistry dataset describes all the moral features of posts that come from the Reddit thread "Am I an asshole". In particular, each post is associated with a list of dyads. Each of this dyad contains the polarization of a specific moral foundation (e.g. care-harm), according to the judgment given to the post.
For the explaining annotation task the most useful features of the Social Chemistry dataset are the following:

- situation: the title of the post
- rot-moral-foundations: the list of moral dyads
- rot-judgment: the judgment given to the action described in the post.

For complexity reason only a small amount of posts have been selected. The dataset has been filtered by selecting those judgements that contains "it's bad". Then the Haidt values have been selected from dyads, according to the judgment.
Separately, the situation column has been lemmatized and filtered to a specific range of text length. Eventually the Social Chemistry dataset has been reduced from three-hundred thousand posts to three thousand.
Once the dataset has been filtered, for each situation the relative RDF is downloaded through FRED [5], synchronized with Framester [6], and extended with ValueNet, in order to achieve the value trigger. This trigger is the semantic element in the RDF that activated the Haidt value, see next section. However, the actual posts which have been analyzed are less than three thousand, due to the incapacity of FRED to build the RDF and the void responses from ValueNet.

III. Analysis

Social Chemistry dataset contains the Haidt value dyads, thus we have computed the distribution of the dyads. Out of the 2929 social chemistry sentences taken in consideration, 1206 have been filtered and not taken in consideration during the analysis, due to no response from FRED and the Value Detector (ValueNet). The distribution of the dyads shows that Harm/Care is the most preponderant dyad in the data (see Table I).
Social Chemistry rot-judgment column gives a positive or negative connotation on the Social Chemistry dyads. In particular, by focusing on "it's bad", only the negative connotation of dyads has been considered, such as Harm in the case of Care-Harm, and the same has been applied for the other dyads.
There are only five Haidt values, encountered in the Social Chemistry dataset: betrayal, cheating, degradation, harm and subversion. On the other hand the Value Detector that extends the RDFs turtles given by FRED, can give as responses the whole set of Haidt Values: betrayal, cheating, degradation, harm, subversion, authority, care, fairness, liberty, loyalty and oppression. Thus, two different accuracy between the results given by ValueNet and the ground truth given by Social Chemistry has been calculated. One accuracy takes in consideration the dyads pair of Social Chemistry and ValueNet responses (see Table II), and the other accuracy takes in consideration the the negative connotation of Social Chemistry data and ValueNet responses (see Table III). The accuracy by considering the pairs is 47.6%, while the other accuracy is 20.9%.

The accuracy of ValueNet is not so high for several reasons. Firstly, the project has been mixing an automated reasoning as in ValueNet with a manual reasoning as what has been done with the Social Chemistry dataset analysis. Secondly, the manual reasoning realized for social chemistry didn't avoid syntactic and semantic errors. Lastly, except for Care-Harm, there is a general lack of representation.

Haidt Value dyad	Distribution
Loyalty-Betrayal	286
Fairness-Cheating	303
Sanctity-Degradation	140
Care-Harm	823
Authority-Subversion	171
Total	1723

TABLE I. SOCIAL CHEMISTRY DYADS DISTRIBUTION

Haidt Value	TP	Accuracy
Betrayal	6	0.3%
Cheating	30	1.7%
Degradation	4	0.2%
Harm	435	25.2%
Subversion	1	0.1%
Loyalty	43	2.5%
Fairness	56	3.3%
Sanctity	0	0%
Care	238	13.8%
Authority	9	0.5%
Total	822	47.6%

TABLE II. VALUENET PERFORMANCE EVALUATION CONSIDERING EACH HAIDT VALUES OF DYADS AS POSSIBLE GROUND TRUTH

Haidt Value	Support	Accuracy	Precision	Recall	F1-score
Betrayal	376		22.2%	1.6%	2.9%
Cheating	396		35.7%	7.6%	12.5%
Degradation	178		15.4%	2.2%	3.9%
Harm	1111		54.6%	39.1%	45.6%
Subversion	219		2.8%	0.4%	0.8%
Total	2280	20.9%	37.9%	20.9%	25.3%

TABLE III. VALUENET PERFORMANCE EVALUATION CONSIDERING ONLY THE NEGATIVE CONNOTATION OF HAIDT VALUES

However there is a relevant issue in Table III. The evaluation metrics are not aligned with the actual predictions. These metrics have been performed by splitting Social Chemistry predictions and ValueNet predictions. In this way it is possible to build the the confusion matrix in Figure I. However splitting the ValueNet predictions makes results not aligned. In particular, if ValueNet predictions are splitted, the relative ground truth is repeated. Thus, many false negative cases will be added into the predictions count. For instance, considering a Reddit post that Social Chemistry has associated with two possible dyads: ”Care-Harm” and ”Loyalty-Betrayal”. Then, according to the negative connotations, the relative Haidt values are

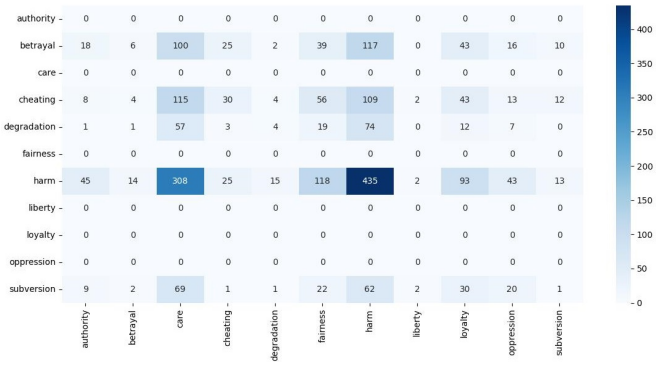

Fig. 1. Confusion Matrix between Social Chemistry negative connoted ground truth and ValueNet predictions.

extracted by these dyads: Harm and Betrayal. When ValueNet analyzes the Reddit post, assuming its predictions are Harm and Degradation, there are 4 predictions (see Table IV), but only one of them is the correct one. So the accuracy is 25%. However Care and Harm are doubled as predictions. Strictly speaking, during the ValueNet performance evaluation, the correct accuracy for this example should be 50%, taking in consideration Harm as the correct prediction and Degradation as the incorrect one. Even though the evaluation of ValueNet for negative connotation is not right, the confusion matrix in Figure 1 is still useful, because it describes the distribution of predictions.

Label true	Label predicted
Harm	Harm
Harm	Degradation
Care	Harm
Care	Degradation

TABLE IV

Analyzing the confusion matrix in figure 1), the Value detector tends to distribute the Haidt values more frequently respect to Social Chemistry. For example, 1/3 of predictions that have been categorized as Care-Harm in Social Chemistry (the Harm row in the confusion matrix) have been categorized in a different way in ValueNet (see Table V).

Dyads	Error Rate
Loyalty-Betrayal	86.9%
Fairness-Cheating	78.3%
Sanctity-Degradation	97.7%
Care-Harm	33.1%
Authority-Subversion	95.4%

TABLE V. VALUENET ERROR RATE.

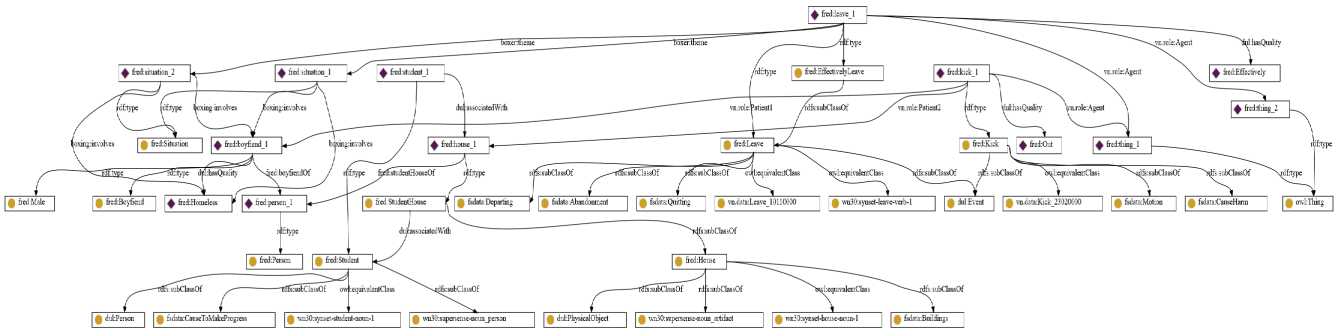

Fig. 2. FRED response

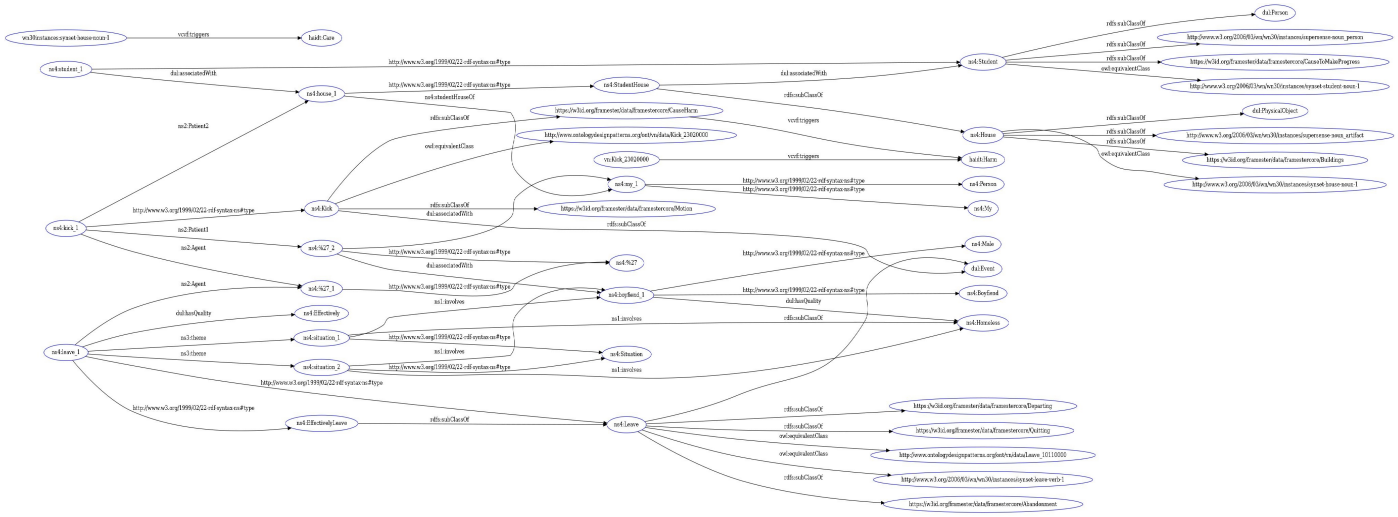

Fig. 3. ValueNet response

Considering the table VI, the majority of the triggers comes from WordNet. However it hasn't been possible to perform any ulterior analysis on WordNet triggers and VerbeNet triggers, due to the fact that they are isolated in ValueNet RDFs. The only triggers that are been classified as frames are the ones which contain some interesting patterns that involve also VerbNet entities. In particulare frame triggers are related to VerbNet roles. So the analysis has been focused on them and the relation between VerbNet roles and value triggers (frames). In other words, understanding the path between these entities on RDFs is a good intuition to understand value annotations.

	Triggers
Wordnet	1272
Verbnet	351
Frame	903

TABLE VI. VALUE TRIGGERS CLASSIFICATION

For instance, let's consider the following sentence:

"Kicking my boyfriend out of my student house and effectively leave him homeless."

This sentence has been converted into an RDF graph, thanks to FRED (see figure 2) and eventually extended by the Value Detector into a new RDF graph (see Figure 3). As can be seen, the sentence has three triggers:

- "wn30istances:synset-house-noun-1" is the WordNet trigger and it is completely isolated
- "vn:kick-23020000" is the VerbNet trigger and it has no enetering arcs on it
- "https://w3id.org/framester/data/framestercore/CauseHarm" is the frame trigger and it is related to the VerbNet rol

The only pattern that has been found connects VerbNet roles and the value triggers with a sub-class schema. In terms of triples, there is a path between the role type subject (or object) and the value trigger, according to the following SPARQL query.

Therefore the value trigger is a super class of the subject's type or of the object's type.

Considering the RDF returned by ValueNet for the previous example (see Figure 3), there is the following path:

```
ns4:kick_1 ns2:Agent ns4:%27_1
ns4:kick_1 a ns4:Kick
ns4:Kick rdfs:subClassOf https://w3id.org/framester/
    data/framestercore/CauseHarm
https://w3id.org/framester/data/framestercore/
    CauseHarm vcvf:triggers haidt:Harm
```

In this example some prefixes are named as "ns" (aka non-specified). The prefix of VerbNet roles (http://www.ontologydesignpatterns.org/ont/vn/abox/role/) is one of these. Moreover, according to the order of the identification provided by FRED, some prefixes can change their "non-specified index" across RDFs. So for these reason in this example "ns2" is the prefix for VerbNet roles. Anyways, in this example the path starts from the type of subject related to the VerbNet role called Agent. This node is "ns4:Kick" and it is linked to the value trigger "CauseHarm" with a sub-class schema.

In general, for each role there is a variable number of internal nodes, in which cycles has been avoided (see Table VII and Table VIII). So the following query has been used to acquire paths, such as the previous example, with different number of sub-class that represent internal nodes.

```
SELECT ?type_sub ?el_0 ... ?el_n ?haidt
WHERE {
    ?sub ns1:role ?obj.
    ?sub a ?type_sub.
    ?type_sub rdfs:subClassOf ?el_0.
    ...
    ?el_n vcvf:triggers ?haidt
}
```

Role	Internal Nodes			
	0	1	2	3
Agent	0	93	2	0
Actor	0	2	0	0
Beneficiary	0	0	0	0
Cause	0	6	0	0
Experiencer	0	4	0	0
Patient	0	0	0	0
Recipient	0	1	0	0
Theme	0	6	1	0
Toward	0	0	0	0

TABLE VII. PATH LENGTH ANALYSIS STARTING FROM THE TYPE OF ROLE SUBJECT

Role	Internal Nodes			
	0	1	2	3
Agent	0	29	3	1
Actor	0	0	0	0
Beneficiary	0	0	0	0
Cause	0	0	0	0
Experiencer	0	0	0	0
Patient	0	13	1	0
Recipient	0	0	0	0
Theme	0	40	2	0
Toward	0	0	0	0

TABLE VIII. PATH LENGTH ANALYSIS STARTING FROM THE TYPE OF ROLE OBJECT

Role	Haidt value	Role Start	Path
Actor	Harm	sub	1
"	Loyalty	sub	1
"	Authority	obj	1
Agent	Betrayal	sub	6
"	Care	obj	6
"	"	sub	15
"	Fairness	obj	2
"	"	sub	6
"	Harm	obj	20
"	"	sub	57
"	Loyalty	obj	2
"	"	sub	7
"	Oppression	obj	2
"	"	sub	4
Cause	Care	sub	1
"	Harm	sub	4
"	Oppression	sub	1
Experiencer	Fairness	sub	2
"	Harm	sub	2
Patient	Care	obj	4
"	Fairness	obj	3
"	Harm	obj	6
"	Oppression	obj	1
Recipient	Care	sub	1
Theme	Betrayal	obj	2
"	Care	obj	6
"	"	sub	1
"	Fairness	obj	2
"	Harm	obj	26
"	"	sub	6
"	Loyalty	obj	2
"	Oppression	obj	4

TABLE IX. PATH COUNT GIVEN ROLE, HAIDT VALUE AND SUB/OBJ TYPE

The next step of the analysis is focused on the count of paths, for each combination of the result feature. According to Table IX, Agent and Theme are the roles which activate paths more than any others, especially from both their subject and object. Furthermore, for each VerbNet role, the Haidt value Harm is always triggered, according to the fact that it's the most triggered Haidt value (see Table I). Consequentially, Harm is strongly correlated with Agent and Theme, also because they produce its major paths.

Lastly, the embedding space of Haidt values has been plotted, considering the paths' first nodes. It has been produced by approximating the multi-dimensional Glove embeddings with PCA. According to the previous example, the node "ns4:Kick" has been converted in a GloVe embedding by removing the prefix.

Even though the embedding space has been plotted, it's hard to understand the behaviour of Haidt values. In Figure 4 each Haidt value covers more or less the same embedding space. However it's possible to infer some information when filtered. For instance, considering Figure 6 different regions can be spotted, one dense and concentrated in the center, representing Betrayal Haidt value, and one sparse for the Fairness one.

It's plausible that these embedding spaces, consequently the paths they represent, will be placed approximately in the same region. Thus, a path that is represented in those areas will be probably classified with the relative Haidt value. However this is only a speculation, due to several issues:

- Except for Harm and Care, there's a lack of representation for the rest of Haidt values.
- Care and Harm are too sparse in the embedding space, see Figure 6.
- Words that belong an Haidt values are not necessary related to it.

The two most representative Haidt values, Care and Harm, are sparse in the embedding space. They are possibly a super-class for the other Haidt values. If so, each Haidt values is semantically related to Care and Harm. This hypothesis is emphasized by the confusion matrix in Figure 1, because Care and Harm are always present in each Haidt value that produce some predictions. Another possibility is that they could be too generic to be placed in a specific region in the embedding space and thus, they are sparse, making difficult any inferring or any consideration on their behaviour.

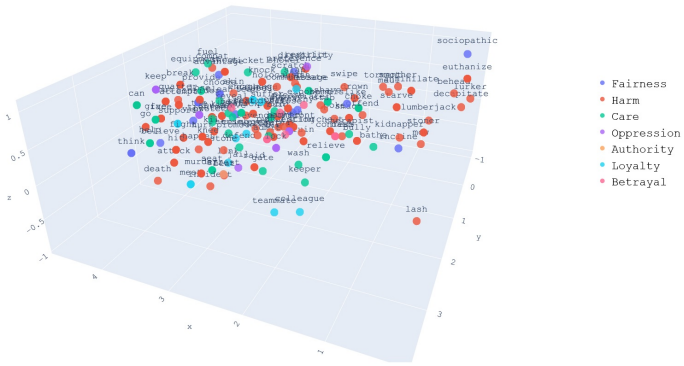

Fig. 4. Embedding space plot of paths start node, according to the specific Haidt value.

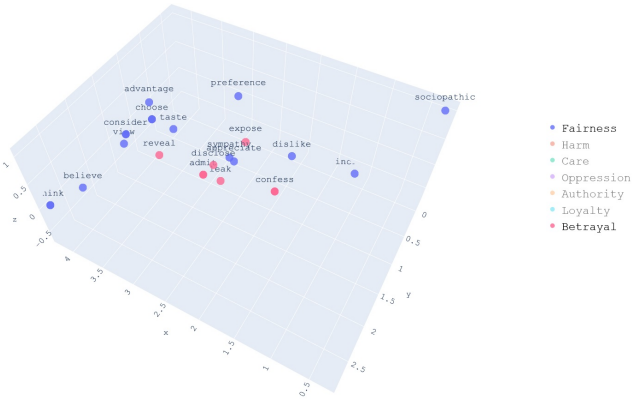

Fig. 5. Embedding space plot for Fairness and Betrayal.

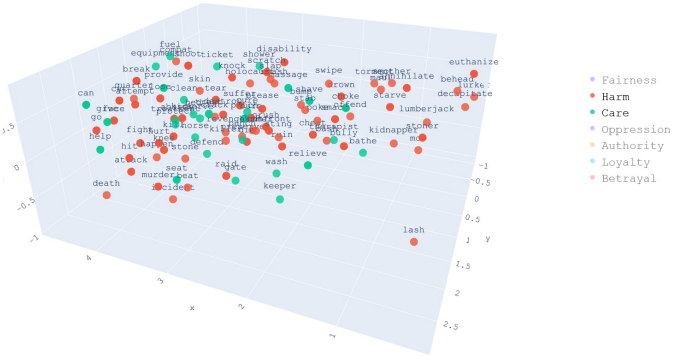

Fig. 6. Embedding space plot of Care and Harm

IV. Conclusion

This paper has started by Reddit posts titles and their dyads that have been acquired by Social Chemistry dataset. The correct Haidt values have been extracted from dyads by selecting those ones that are related to a negative connotation, in order to focus on specific kind of moral polarity. Reddit posts titles have been represented as RDFs, thanks to FRED and they have been extended with ValueNet, in order to acquire triggers that activate the corresponding Haidt values. Even though ValueNet has 47.6% of accuracy, its predictions are related to more Haidt values than Social Chemistry ground truths. Considering that the only triggers that provides some extra relation in RDFs was frame triggers, this paper have focused the analysis on those paths in RDFs, provided by the Value Detector, in which the frame trigger was connected to some ontology. It has been discovered that there is a relation between VerbNet roles with the frame triggers and Haidt values. For each path that describes this relationship, its first node has been selected as the path representative and it has been converted into a GloVe embedding and plotted through PCA to produce a 3D embedding space. In this way it is theoretically possible explain the reason why the Reddit posts have been categorized with specific Haidt Values because their paths that relate Haidt values are represented in specific areas in the embedding space. However, the actual result was that every path has been condensed into an area that is delimited by Care and Harm Haidt values embedding space. Considering this results, there are two possible causes. Firstly, the Haidt values Care and Harm can be super-class of the other ones. Secondly they can be too generic and so they are too sparse in the embedding space. However the first hypothesis is highly emphasized by the confusion matrix of the Value Detector's predictions (see Figure 1) which shows that the Haidt values Care and Harm are always present, as misclassifications, in the other predictions, provided by Social Chemistry dataset.

References

[1] *Moral foundation*. https://en.wikipedia.org/wiki/Moral_foundations_theory.

[2] *ValuesFromText*. https : / / github . com / Frankgamer97 / ValuesFromText.

[3] *RDF*. https://it.wikipedia.org/wiki/Resource_Description_Framework.

[4] Vered Shwartz Maarten Sap Yejin Choi Maxwell Forbes Jena D. Hwang. *Social Chemistry*. URL: https : / / maxwellforbes.com/social-chemistry.

[5] *FRED*. http://wit.istc.cnr.it/stlab-tools/fred/demo/.

[6] University of Bologna. *Framester*. URL: http://etna.istc.cnr.it/framester_web/.