From Stance to Concern: Adaptation of Propositional Analysis to New Tasks and Domains

Anonymous ACL submission

Abstract

We present a generalized paradigm for adaptation of propositional analysis (predicate-argument pairs) to new tasks and domains, leveraging an analogy between *stances* (belief-driven sentiment) and *concerns* (topical issues with moral dimensions/endorsements). A key contribution is the combination of semi-automatic resource building for extraction of domain-dependent concern types (with 2-4 hours of human labor per domain) and an entirely automatic procedure for extraction of domain-independent moral dimensions and endorsement values. Prudent (automatic) selection of terms from propositional structures for lexical expansion (via semantic similarity) produces new moral dimension lexicons at three levels of granularity beyond a strong baseline lexicon. We develop a ground truth (GT) based on expert annotators and compare our concern detection output to GT, to yield 231% improvement in recall over baseline, with only a 10% loss in precision. F1 yields 66% improvement over baseline and 97.8% of human performance. Moreover, our lexically based approach yields large savings in terms of human labor and costly model building. Work produced herein provides to the community a newly expanded moral dimension/value lexicon, annotation guidelines, and GT.

1 Introduction

This paper presents a generalized paradigm for adaptation of tasks involving predicate-argument pairs, i.e., combinations of actions and their participants, to new tasks and domains. Predicate-argument analysis has been a longstanding area of research for many tasks: event detection (Du and Cardie, 2020; Zhang et al., 2020), opinion extraction (Yang and Cardie, 2013), textual entailment (Stern and Dagan, 2014), and coreference (Shibata and Kurohashi, 2018). We refer to the induction of such representations as *propositional analysis*. We induce a proposition PREDICATE(x_1,x_2,...) to represent sentences such as *John wears a mask*: wear(John,mask). We then pivot off this representation to answer questions such as *What is John's stance towards mask wearing?* or *What concerns does John have about mask wearing?*

Stance detection has recently (re-)emerged as a very active research area, yet many approaches generally equate stance to raw (bag-of-word) sentiment and often employ machine-learning based models requiring large amounts of (labeled or unlabeled) training data. Within such approaches, the notion of *stance* varies, but generally falls into one of a handful of "sentiment-like" categories for *stance holder* X regarding topic Y, i.e., X agrees/disagrees with Y (Umer et al., 2020), X favors/disfavors Y (Krejzl et al., 2017), X is pro/anti Y (Samih and Darwish, 2021), X has a positive/negative opinion about Y (AlDayel and Magdy, 2021), or X is in favor/against/neither Y (Küçük and Can, 2020).

We adopt the stance definition of Mather et al. (2021), originally formulated for Covid-19. There, a stance is a **belief-driven sentiment**, derived via propositional analysis (i.e., *I believe masks do not help [and if that belief were true, I would be anti-mask]*), instead of a bag-of-words lexical matching or embedding approach that produces a basic pro/anti label. This variant of stance detection uses a proposition to identify a domain-relevant belief in the Covid-19 domain; the belief is leveraged to compute a belief-driven sentiment and attitude toward a topic in that domain (e.g., *masks*). For example, *I believe masks do not protect me* is rendered as a belief type PROTECT with an underlying proposition protect(masks) and a negative overall stance toward the propositional content: *masks*.

We implement and evaluate an analogous propositional framework for a new task, *concern detection*, in the new domain of the 2017 French elections. Table 1 shows an output for both stance and concern detection. The proposition common to both is ruin(jean-luc melenchon, economy). For

Stance Proposition: *ruin(jean-luc melenchon, economy)*	
Belief Type	**Belief/Sent Values**
DESTROY	belief strength: 2.5, sentiment strength: -1

Concern Proposition: *ruin(jean-luc melenchon, economy)*	
Concern Type	**Moral Dims/Values**
ECONOMIC	care: 1.4, purity: 2.75

Table 1: Representative Stance and Concern system output for a given tweet in the 2017 French Election Domain: *Marine Le Pen LEADS in French poll as far left Jean-Luc Melenchon 'could ruin economy'*. Belief ranges from -3 to 3, sentiment from -1 to 1, and the moral dimensions from 1 to 9.

stance, a belief *type*, DESTROY, is coupled with *values*: (1) Belief strength captures a belief commitment ranging from certainty that the belief is not true (-3) to certainty that the belief is true (+3), with 0 as uncommitted belief; and (2) Belief-driven sentiment strength ranges from extremely negative (-1) to extremely positive (+1), with 0 as neutral.

We define a *concern* to be a topical *type* (e.g., ECONOMIC) coupled with a set of *values*: (1) moral dimensions from Moral Foundations Theory (MFT) (Haidt and Joseph, 2004; Graham et al., 2009, 2011, 2012), represented as vice/virtue pairings (authority/subversion, care/harm, fairness/cheating, loyalty/betrayal, and purity/degradation); and (2) corresponding endorsement values, where "vice" is between 1 to 5 and "virtue" is >5 to 9, with 1 being strong vice and 9 being strong virtue. Dimensions and endorsements shown in Table 1 are derived from a state-of-the-art baseline (Moral 1), see §3.

We demonstrate that the key to reduced adaptation time is the coupling of semi-automatic resource building for concern types with automated expansion of domain-independent concern values using semantic similarity. We develop a ground truth (GT) based on expert annotators and compare concern detection output to GT. We also demonstrate that, with each lexicon expansion, the performance of concern detection improves significantly over a state-of-the-art baseline moral dimension lexicon. We obtain a 231% improvement in recall over a strong baseline for our best performing system, with only a 10% loss in precision.

2 Background and Motivation

This section provides background and motivation for task and domain adaptability applied to concern detection in the 2017 French Election domain (§2.1, §2.2), including concern values induced from Moral Foundations Theory (MFT) (§2.3).

2.1 Task and Domain Adaptability

Task adaptability and domain adaptability are two supporting areas of research for this work. Prior domain adaptation approaches, surveyed by Ramponi and Plank (2020), have been applied to tasks such as sentiment analysis (Ben-David et al., 2020; Ghosal et al., 2020), stance detection (Xu et al., 2019), and event trigger identification (Naik and Rose, 2020). Task adaptation approaches (Gururangan et al., 2020; Garg et al., 2020; Ziser and Reichart, 2019) have been applied to tasks such as answer selection for question answering.

To date, both types of adaptation rely heavily on machine learning (ML) techniques, many of which require a large amount (e.g., 1M+, Gururangan et al. (2020)) of training data (whether labelled or not). Some approaches employ smaller datasets, e.g., 10K+ Amazon reviews, fake news articles (Ben-David et al., 2020; Xu et al., 2019). Additionally, while explainability has recently been brought to the fore in deep learning approaches, as surveyed by Xie et al. (2020), such systems have not focused on task and domain adaptability.

We develop a general framework for resource building techniques that features task adaptability and retains the ability to adapt quickly to new domains. Our dataset requirements are much more minimal than prior approaches (2500 tweets per domain), there is no human labeling of corpora, and no model training required. Moreover, explainability is achieved by virtue of inclusion of propositional information (who did what to whom) that serves as a window into the process of detecting concern types and moral dimensions.

2.2 Stance and Concerns

Pirolli et al. (2021) applied a belief-based formulation of stance in the Covid-19 domain, with topics such as *mask wearing* and *social distancing*. For example, a *stance* assigned to *Wear a mask!* includes a PROTECT belief *type*, where the predicate *wear* is considered a "trigger" and *a mask* is considered the "content" of the belief. The *values* associated with this stance include a belief strength of

+3 and a sentiment strength of +1. The final *stance* is thus a belief-oriented sentiment with this interpretation: the person posting the tweet is positive toward "masks," assuming the belief that masks are protective is true. In the 2017 French Election domain, a stance representation (e.g., for the example in Table 1) would be: <DESTROY(*ruin(jean-luc melenchon, economy)*), Bel:+2.5, Sent:-1>.

While this prior framework lays the groundwork for domain adaptability, it has not been shown to be generalizable to new tasks (within or across domains), which is the focus of this paper. We leverage the propositional underpinnings of the framework of Mather et al. (2021) to enable a straightforward adaptation from stance detection to a new task, while also retaining domain adaptability. The new task of interest is *concern detection*. This task involves extraction of a *concern type* (e.g., *immigration, taxation*) associated with a given domain (e.g., *2017 French Elections*), analogous to the extraction of a *belief type* for a given stance detection domain. In this paper, we demonstrate that it is straightforward to port belief-targeted stance both to a new domain (French elections) and, through an analogous proposition-based extraction, to a new task: Concern detection. An example of a formal Concern representation in the 2017 French Election domain for the example in Table 1 would be: <ECONOMIC(*ruin(jean-luc melenchon, economy)*), Care: 1.4, Purity: 2.75>.

The approach described herein focuses on lexicon expansion obtained automatically through semantic similarity to map key terms in propositional statements to moral foundation lexicon words, using WordNet (Fellbaum, 1998). Three different variants of lexicon expansion (described in §3.3) improve on results obtained using the current state-of-the-art moral lexicon of Araque et al. (2020), which we take to be a strong baseline (henceforth referred to as 'Moral 1'). The advance beyond this prior work lies in the prudent (automatic) selection of terms designated for expansion, based on propositional structure, and the combination of moral dimensions with concern types.

2.3 The MFT Framework and Influence

We focus on concern detection because identifying what people are discussing online within a particular domain is important and useful, as is identifying the moral justifications or deliberate appeals to moral identity in these discussions. We use the Moral Foundations Theory (MFT) framework (Haidt and Joseph, 2004; Graham et al., 2009, 2011, 2012) to encode the moral dimensions of social media contributions. These moral dimensions may serve as potential indicators of influence attempts, as in *When it comes to immigration it's not about children, it's about damaging our country!*, where the Concern type is IMMIGRATION_REFUGEE and there is an appeal to the vice side (*harm*) of the *care/harm* moral dimension.

An emphasis on highly controversial and/or polarizing topics in online posts/messages may be indicative of an attempt to sway others. More importantly, if these posts/messages are interwoven with language that reflects (and speaks to) the moral values of the target audience it can increase in-group cohesion, and that may further contribute to polarization. Additionally, deliberate use of morality to justify harmful intentions towards others may foster online outrage disguised as ethical conduct (Bandura et al., 1996; Friedman et al., 2021).

Several studies have shown that social groups provide a framework in which moral values are endorsed, and when these values are threatened by e.g., opposing political ideology, existing beliefs of the group are strengthened (Van Bavel and Pereira, 2018). Thus, when presented with information that is incongruent with our identity and in-group, we tend to override accuracy motives in favor of social identity goals (partisan bias). These studies have shown that, when accuracy and identity goals are in conflict, moral values determine which belief to endorse and thus how to engage with information. This makes moral values an ideal breeding ground for influence campaigns, on a small scale (e.g. political ideology) as well as on larger scale (e.g. cultures or subcultures), but also very useful for our concern detection model.

We note that most studies of cross-cultural values, beliefs, and morality have been conducted by WEIRD (white, educated, industrialized, rich, and democratic) countries (Goodwin et al., 2020; Henrich et al., 2010). Inglehart's model of World Values (Inglehart and Welzel, 2010) has surveyed 60 countries for the last 40 years, taking into account that many nations are more concerned with economic and physical security (e.g. survival), while self-expression values are more reflective of Western countries. For example, in Pakistan or Nigeria 90% of the population say that God is extremely important in their lives, while in Japan only 6%

take this position. Similarly, Schwartz' Theory of Basic Values (Schwartz, 2012) uses a different set of organizing principles, e.g., values that relate to anxiety (e.g., tradition, security, control of threat) which may lead to an increased belief in misinformation (Jost et al., 2003). Thus, moral dimensions combined with concern types might be a potential indicator of a common actor (possibly an outside influencer) if several individuals or accounts (potentially purporting to be individuals) invoke the same moral dimensions across their messages.

3 Task and Domain Adaptation

Adaptation of stance detection to concern detection gives rise to a new framework for rapid development of a task-adapted system that retains domain adaptability and uses relatively low amounts of data, with only 2-4 hours of human categorization.

3.1 Resource Building Generalizations

We generalize to a new task while retaining domain adaptability by leveraging the stance-concern analogy, through: (a) semi-automatic domain-dependent extraction of *types* from propositional analysis, i.e., moving from *belief types* for stances to *concern types* for concerns; and (b) fully automatic domain-independent induction of associated *values* from a combination of propositional arguments and lexical and semantic resources, i.e., belief/sentiment *strengths* (cf. (Baker et al., 2012)) for stances and moral dimensions and endorsements (cf. (Graham et al., 2012)) for concerns.

Resource building for domain-dependent stance *types* involves propositional analysis using semantic role labeling (SRL) (Gardner et al., 2018) to detect positions with the most highly relevant content terms, e.g., *masks*. The work of Mather et al. (2021) indicates that these positions are ARG0 and ARG1. To port this approach over to the induction of domain-specific concern *types*, we conducted a similar analysis and found that the same positions (ARG0 and ARG1) contain the most highly relevant terms for concerns, e.g., *economy*. Semi-automatic induction of concern types thus leverages these positions, as described in §3.2.

Just as stance resource building induces domain-independent stance *values* (belief / sentiment strengths), concern resource building induces domain-independent concern *values* (moral dimensions / endorsements). A deeper propositional analysis reveals that additional SRL positions have a high likelihood of association with moral dimension terms, e.g., *ruin*: V, ARG2, ARGM-ADV, ARM-MR, ARGM-PRD. This discovery further generalizes the original stance resource-building approach and enables rapid task adaptation to concerns through entirely automatic means. We leverage these additional SRL positions to extract candidate terms for expansion of moral dimensions. Associated endorsements are then inherited from semantically similar terms from baseline Moral 1. Further details about the expansion of moral dimensions are provided in §3.3.

Domain adaptability is retained—on analogy with stance detection—by separating and independently addressing two aspects of concern detection: (a) induction of domain-specific concern types; (b) induction of domain-independent moral dimensions. Lexicon expansion using this approach can thus be applied to domains beyond the French Elections presented herein.

3.2 Concern Type Induction

We adopt a generalized semi-automatic process for concern type induction that retains domain adaptability. A small set (approximately 15) of domain-relevant key terms (e.g., health, taxation, immigration), is provided by the domain expert as input to the semi-automatic resource building tool. These terms are used to filter the domain relevant dataset (an English subset of a Kaggle Twitter dataset for the 2017 French Elections (Daignan, 2017) of 1.9 million tweets). The filtered tweet subset (214k) is then divided into training (2500),[1] and development (211,500) subsets, and propositional analysis is applied to the training set in a 3-step process.

First, the top 25 most frequent terms are extracted (e.g., *economy)*,[2] ignoring functional elements such as stop words. Second, the verbs whose relevant SRL positions (defined in §3.1) contain any of these top 25 terms are extracted, and the top 10 most frequent verbs e.g., *ruin, restrict* are selected for further processing. Lastly, the top 40 most frequent terms associated with relevant SRL positions (for each of the 10 verbs) are extracted, e.g., *economy and business*, yielding 400 propositions, e.g., *ruin(economy), restrict(business)*.

The propositions, coupled with terms from the original domain-relevant key terms, are presented

[1] Training data are strictly for one-time semi-automatic resource building, not for model training.

[2] spaCy (Honnibal et al., 2020) is used for preprocessing (e.g. sentence splitting, POS tagging).

to a domain expert, who leverages this input to create concern types (approx. 10 or fewer) and categorizes the terms according to those types. Terms left uncategorized by the expert are dropped. This semi-automated concern-type induction takes 2-4 hours owing to the extraction of high frequency domain relevant propositions.

Concern types induced for the French Election domain are: IMMIGRATION_REFUGEE, ELECTORAL_PROCESS_VOTING_LAWS, ENVIRONMENT_CLIMATE_CHANGE, HEALTH_CARE, TAXATION, ECONOMIC, SOCIAL_SERVICES, INTERNATIONAL_TRADE, MILITARY_ENGAGEMENT, CRIMINAL_JUSTICE.

3.3 Concern Value Induction

Concern *values* leverage MFT (Haidt and Joseph, 2004; Graham et al., 2009, 2011, 2012) and particularly the moralstrength library (Araque et al., 2020), which serves as a strong baseline (referred to as 'Moral 1').[3] Our approach transcends this earlier paradigm through the application of propositional analysis with semantic-role labeling (SRL Gardner et al. (2018)), coupled with a more in-depth WordNet (WN Fellbaum (1998)) expansion, yielding higher recall while more linguistically constrained to retain acceptable precision.

Expansion of moral dimensions relies on propositional analysis, SRL, and WordNet expansion. We select candidates for moral dimension expansion through extraction from propositional statements, and then induce three lexicon variants (in addition to the Moral 1 baseline, and a Moral 0 random chooser described in §6) using a progression of finer-grained WordNet-based semantic-similarity functions. This expansion supports the goal of task adaptability, as required in the transition from stance detection to concern detection.

The end result is a general approach to induction of *values* for specific tasks, rendered in the form of domain-independent lexicons. That is, analogous to belief/sentiment terms of stance detection (*might*, *probably*, *hate*, *love*), we induce vice/virtue moral dimensions and their corresponding endorsement values for concern detection.

This expansion results in three different system variations for moral dimension/values (referred to as Moral 2, Moral 3, Moral 4) beyond the Moral 1 baseline. We note that moral dimensions are assigned automatically to each lexical entry via semantic similarity to Moral 1 terms; the endorsement values are then *inherited* from the most semantically similar term from the Moral 1 lexicon. An excerpt of lexicon output is shown below, from the best performing lexicon (Moral 4):

hypocrite - dim: betrayal; endorse: 1.0 (strong)
appreciation - dim: care; endorse: 8.57 (strong)

Lexicon Expansion Details: Moral 2-4 rely on automatic propositional analysis for prudent selection of words from the training data (via SRL, see §3.1) to be considered candidates for lexicon expansion. The highest similarity match is recorded to inherit the corresponding moral endorsement value from Moral 1. A brief description of all lexicon expansions used for induction of moral dimensions and values for concern detection is provided below. (See detailed description in Appendix A and links to Moral 2-4 lexicons in Appendix B.)

Note: *The term "lemma-matched" below refers to a match between a word in a training tweet and a synset's first lemma.*

Moral 1: This initial moral dimension lexicon developed by Araque et al. (2020) serves as a strong baseline, with 2800+ terms across five moral dimensions. This was derived by expanding an initial crowd sourced lexicon of about 480+ terms, annotated for moral dimensions and endorsements. Expansion to 2800+ terms was via WordNet synset matching, without regard to propositional analysis.

Moral 2: (Added 214 terms, for total of 3064) This lexicon expansion yields a set of terms whose lemma-matched synsets are semantically similar (above a threshold) to lemma-matched synsets of the words in the strong baseline (Moral 1) lexicon.

Moral 3: (Added 995 terms, for total of 3845) This lexicon expansion yields a set of terms whose lemma-matched synsets and their descendents are semantically similar (above a threshold) to lemma-matched synsets and their descendents of the words in the strong baseline (Moral 1) lexicon.

Moral 4: (Added 5623 terms, for total of 8473) This lexicon expansion yields a set of terms drawn from **all** synsets and their descendents that are semantically similar (above a threshold) to **all** synsets and their descendents of words in the strong baseline (Moral 1) lexicon, without lemma matching.

4 Annotation for Ground Truth

We conduct an annotation task to develop Ground Truth (GT), against which to compare our concern detection system variants, based on 50 held-out

[3] https://github.com/oaraque/moral-foundations.

tweets from the held-out development portion of the English subset of the Kaggle Twitter dataset on 2017 French elections (Daignan, 2017). Annotation was completed by two non-algorithm developers (one with expertise in linguistics, the other with expertise in psychosocial moral indicators). Annotators were provided guidelines (Appendix C) for both concern types and moral dimensions. For concern types the inter-rater reliability (IRR) is calculated through macroaveraging of kappa scores (Carletta, 1996) which produces a 66% agreement, considered **Strong** according to (McHugh, 2012). By contrast, the macroaveraged IRR for moral dimensions is low (16%), which is considered **Weak**.

Given the reliability of the concern type annotations, our system output is compared against the union of both annotators to the concern type scores shown 3. However, for moral dimensions, our system output is compared only against the single annotator with expertise in psychosocial moral indicators. We consider this a reasonable choice given that research in detecting moral dimensions is still in nascent stages and significant training for the task is required to achieve a reliable GT. Furthermore, the notion of "ground truth" is inherently problematic with moral indicators given the complex way in which this dimension varies with socio-cultural factors.

The Ground Truth resulting from the annotation described above is in Appendix D. Data used herein (Daignan, 2017) is an open and publicly available dataset carrying no privacy or copyright restrictions. Furthermore, IRB formally designates this research as non human subject research. Annotators spent two hours apiece on the task.

5 Sample Runs

We have implemented and validated a system to detect concerns that relies on induced lexical resources. Using spaCy and SRL we derive a propositional representation coupled with a concern *type* and a set of concern *values* (moral dimensions/endorsements). System run time per tweet is, on average, 0.6s, using Python on a MacBook Pro.

The system has been run on a 1K-tweet held-out portion of the development dataset, with representative outputs shown in Table 2. These are taken from our best performing variant described in §6. In Example 1, the IMMIGRATION_REFUGEE Concern *type* is induced by *refugees*, coupled with appropriate moral dimension/endorsement *values*:

Example 1: *Vatican's completely surrounded by a wall w/an entrance that's guarded 24/7 & refugees r not allowed* **Concern type**: IMMIGRATION_REFUGEE **Proposition**: *surrounded(vatican, by a wall ...guarded 24/7 & refugees r not allowed))* **Dimensions & Endorsements**: Harm: 3.46, Authority: 6.02, Degradation: 2.71
Example 2: *Where is the justice?* **Concern type**: CRIMINAL_JUSTICE **Proposition**: *is(where, the justice)* **Dimensions & Endorsements**: Fairness: 7.6
Example 3: *Most Melanchon voters care more about their country than their ideological purity.* **Concern type**: ELECTORAL_PROCESS_VOTING_LAWS **Proposition**: *care(most melanchon voters, more about their country)* **Dimensions & Endorsements**: Care: 8.80, Loyalty: 7.14, Authority: 5.57, Degradation: 3.69

Table 2: Sample Output from Concern Detection

Harm (*wall, entrance, guard*), Authority (*wall, guard*), and Degradation (*entrance*). In example 2, the CRIMINAL_JUSTICE Concern *type* is induced (*justice*), coupled with appropriate moral dimension/endorsement *values*: Fairness (*justice*). Both of these appear to be reasonable outputs.

In example 3, the ELECTORAL_PROCESS_VOTING_LAWS Concern *type* is induced (*voters*), coupled with moral dimension/endorsement *values*: Care (*care*), Loyalty (*country*), Authority (*care, country*), Degradation (*care, voters*). Note that this example overgenerated, inducing the Degradation dimension from the terms *care and voters*. This is an indication that further lexicon enhancement is needed (as alluded to in §7) for elimination of spurious lexical entries that potentially lead to false positives (i.e., a reduction in Precision).

6 Results and Analysis

We compare concern detection output using 4 different variants of moral dimension lexicons. Our (strong) baseline is "Moral 1" (Araque et al., 2020). Subsequent variants (Moral 2-4) are the expanded moral dimension lexicons using the techniques described in §3.3. The results of each system output are compared to Ground Truth (GT) as shown in Table 3.

System performance is measured by weighted macro-averaged precision (P), recall (R), and F1 scores. Domain-dependent Concern *Type* is not affected by moral lexicon variants and independently has its own P/R/F1 scores. Concern *Values* (i.e., moral Dimensions) are impacted by lexicon variants and therefore have a row corresponding to each variant. We note the importance of applying a **weighted** macro average to these scores due an imbalance in the distribution of classes (Delgado and Tibau, 2019), where the probability of one class can be substantially higher or lower from others. For example, we observed that the number of ELECTORAL_PROCESS_VOTING_LAWS annotations is 2.3 times higher than the number of INTERNATIONAL_TRADE annotations.

We also provide a system comparison against a random chooser for Concern Types (Concern 0) and Moral Values (Moral 0). At first glance this random chooser would appear to have a decent likelihood of getting many hits, with an expected rate of about 50% that each Concern type and Moral value will be selected. If so, this would result in an expected 250 positive selections for concern types (whereas the two annotators together only made 60 positive selections) and an expected 250 positive selections for moral values (whereas the expert annotator only made 66 positive selections). However, the results in Table 3 indicate that, while the increased number of hits leads to a reasonably high recall, the number of false positives (233) swamps out the hit rate, leading to a low F1 score (20.54). Accordingly, Concern 1 easily beats the random choice baseline by a healthy margin, with an F1 score of 77.17.

In contrast, Moral 0 achieves 79.65% of the performance of the Moral 1 baseline, with an F1 score of 19.65—not too far off from the 24.67 baseline F1 score. Moreover, the F1 scores for Moral 2-4 surpass this baseline, with statistically significant improvements indicated between all system pairs at the 3.5% level or better, according to the McNemar statistical test (McNemar, 1947).[4] That is, all lexicon expansion improvements are statistically significant. Notably, a 231% improvement in Recall is achieved for Moral 4 over the strong baseline (Moral 1): 65.15 vs. 19.69. This is achieved with only a 10% loss of precision, ultimately yielding an F1 score of 40.85 which is a 66% improvement.

[4]Tested values are correct responses (TP or TN) vs. incorrect responses (FP or FN), for determining significance of change in total error rate.

We conduct a further exploration to evaluate the performance of Concern 1 relative to that of human performance. Average F1 score for the two annotators is 78.88, and Concern 1 performance is 77.17 F1. Concern detection thus yields 97.8% of human performance on concern type detection.

Our error analysis of FP/FN cases for concern types reveals that concern detection fails to assign any concern type (FN) to *Infighting among left-wing could hand Front National VICTORY*, which is annotated as ELECTORAL_PROCESS_VOTING_LAWS. This is because terms like *Front* and *left-wing* are not present in the concern type lexicon.

For concern values (moral dimensions), annotator does not assign a moral dimension to *Macron is center right*, yet Moral 4 inaccurately detects (FP) Care, Authority, and Betrayal due to the existence of the word *center* and also detects Purity from the word *right*. For this same sentence Moral 1 also inaccurately detects Fairness from the word *right*. Many cases similar to these impact precision values for each Moral 1-4, potentially requiring lexicon tuning (see §7).

System	TP	FP	FN	P	R	F1
Concern 0	35	233	25	12.69	53.88	20.54
Concern 1	**40**	**1**	**20**	**91.60**	**66.67**	**77.17**
Moral 0	33	225	33	12.74	43.01	19.65
Moral 1	13	**28**	53	**32.99**	19.69	24.67
Moral 2	17	58	49	31.20	25.76	28.22
Moral 3	24	122	42	31.38	36.36	33.69
Moral 4	**43**	181	**23**	29.75	**65.15**	**40.85**

Table 3: Evaluation of Concern Types (Domain-Dependent) and Moral Dimensions (Domain-Independent) computes an unweighted sum of true positives (TP), false positives (FP), and false negatives (FN) across all labels. Precision (P), recall (R), and F1 numbers for all systems are weighted macro-averages except for the random choosers (Concern 0 and Moral 0). Lexicon variants affect only the moral dimensions.

7 Discussion and Future Work

We implement and validate a generalized paradigm for adaptation of tasks involving propositions to a new task (concern detection) and domain (French elections). Our primary contribution is the provision of a framework for rapid adaptability, based on: (a) semi-automatic domain-dependent extraction of *types* from propositional analysis; and (b) fully automatic domain-independent induction of *values* from propositional arguments and semantic

resources. We demonstrate that the coupling of (a) and (b) leads to rapid ramp-up resource construction for a new task and domain (2-4 hours).

We also demonstrate that a deeper propositional analysis is key to generalizing domain-adaptable resource-building for new tasks. We develop a new automatic procedure for expanding moral dimensions (Graham et al., 2012) that incorporates propositional analysis, semantic-role labeling, and in-depth WordNet (WN) expansion, to produce three increasingly expanded lexicons with moral dimensions/endorsements. We develop a ground truth (GT) based on expert annotators and compare our concern detection output to GT, to yield 231% improvement in recall over baseline, with only a 10% loss in precision. F1 yields 66% improvement over baseline and 97.8% of human performance. Moreover, our lexically based approach yields large savings in terms of human labor and costly model building. Work produced herein provides to the community a newly expanded moral dimension/value lexicon, annotation guidelines, and GT for 50 tweets.

Results from system runs have been obtained without any tuning of the lexicons to remove terms that result in a high number of false positives (FP) and false negatives (FN). Future work will explore fine tuning of the lexicons to address cases seen in §6 with an eye toward improving the precision without a large drop in recall, to yield an even higher F1-score.

Another area of future work is incorporation of additional moral dimensions, most notably the liberty/oppression dimension proposed by (Haidt, 2012), in opposition to the authority dimension ('liberal authority' vs. 'attempted domination'). This reflects political equality related to dislike of oppression and concern for victims, not a desire for reciprocity. In political discourse, this is apparent in anti-authoritarianism and anti-government anger, which makes it an important dimension for the topic of the French election (Iyer et al., 2012) and widens the range of potentially relevant information that could indicate influence attempts.

Future work also includes exploration of other cultural frameworks, in addition to or in place of Moral Foundations Theory, e.g. Inglehart's Cultural Map model (Inglehart and Welzel, 2010) and Schwartz Value Theory (Schwartz, 2012). Cultural models that allow more room for non-Western values (e.g. survival needs) are important considerations in order to reduce bias, and a feasible avenue for improving the performance and applicability of concern detection.

Establishing GT for a given cultural context remains difficult, especially when using diverse annotators. Prabhakaran et al. (2021) show that systematic disagreements between annotators (due to socio-cultural backgrounds and/or lived experiences) may be obfuscated through aggregating crowd-sourced annotations. Future work may be improved on, using vector-based approaches assigning weights based on their representativeness in a given culture.

Cultural values are also reflected in language (e.g., gendered vs. non-gendered languages, culturally intrinsic concepts). Accordingly, our future work involves processing concerns for other languages through adaption of SRL (Gardner et al., 2018) to multilingual input, starting with French, employing multilingual preprocessing via spaCy (Honnibal et al., 2020) and EuroWordNet and related multilingual WordNets (Bond and Paik, 2012; Bond and Foster, 2013), to expand moral dimensions to other languages.

Language biases on which some of our work is built on are a potential limitation. For example, cultural idioms like *fluctuat nec mergitur* may be translated from Latin into the correct literal meaning of *tossed [by the waves], but not sunk*, but the culturally distinct values (Paris' coat of arms and motto with a deep affective history) will get lost in translations into English, and with it, its cultural meaning. Similarly, while the English WordNet provides one of the most comprehensive semantic ontology of words in English, embedded biases are still present today, e.g., offensive, racist, and misogynistic slurs (Crawford and Paglen, 2021); such limitations need to be addressed within the language resource community.

Ethics: Risk of misuse of this technology is mitigated by the transparent nature of concern detection, owing to the propositional representations that underlie and inform algorithmic decisions. In contrast to ML approaches, any form of misuse within the technology would be easily discoverable. The technology further serves as a framework within which cultural distinctions may be studied and better understood, thus mitigating the potential for cross-culturally undetected misuse.

A Lexicon Expansion Detailed Descriptions

Below is the detailed description of each expansion algorithm. **Basic definitions:**
w = word in training tweet (e.g., 'concerned')
m = moral foundations word (e.g., 'concern')
wn = wordnet package
S_w = wn.synsets(w) (set of synsets for word w)
$s_{w,1}$ = wn.synsets(w)[0] (first synset for word w)
L_i = lemmas associated with S_i
$l_{i,k}$ = s.lemmas() (lemmas for synset s_i)
$l_{i,1}$ = s.lemmas()[0].name() (first lemma synset s_i)
Note: Use of the term "lemma-matched" below refers to a match between a word in a training tweet and a synset's first lemma.

Lexicon Expansion: All lexicon expansions (Moral 2-4 below) beyond a strong baseline (Moral 1) rely on propositional guidance to select words from the training data to be considered candidates for lexicon expansion. The highest similarity match is recorded for future selection of the moral endorsement value. All moral lexicon expansions (Moral 2-4 below) beyond a strong baseline (Moral 1) apply to each pair (w, m) for each word of interest w and each moral foundations word m ($w \times m$ iterations):

Moral 1: This initial moral dimension lexicon developed by Araque et al. (2020)[5] serves as a strong baseline, with 2800+ terms across five moral dimensions. This was derived by expanding an initial crowd sourced lexicon of about 480+ terms, annotated for moral dimensions and endorsements. Expansion to 2800+ terms was via WordNet synset matching, without regard to propositional analysis.

Moral 2: (Added 214 terms) This lexicon expansion yields a set of terms whose lemma-matched synsets are semantically similar (above a threshold) to lemma-matched synsets of the words in the strong baseline (Moral 1) lexicon, using the following steps: (a) Extract all synsets $s_{w,i}$ in S_w (for word w) whose first lemma $l_{i,1}$ exactly matches w, producing S_x.(b) Extract all $s_{m,j}$ in S_m (for word m) whose first lemma $l_{j,1}$ exactly matches m, producing S_y. (c) Return all lemmas $l_{k,1}$ of any $s_{x,k}$ in S_x that matches any $s_{y,h}$ in S_y using wordnet wup_similarity w/ threshold 0.9 (if any). Example: concerned → concerned.a.01.

Moral 3: (Added 995 terms) This lexicon expansion yields a set of terms whose lemma-matched synsets and their descendents are semantically similar (above a threshold) to lemma-matched synsets and their descendents of the words in the strong baseline (Moral 1) lexicon, using the following steps: (a) Extract all synsets $s_{w,i}$, in S_w (for word w) whose first lemma $l_{i,1}$ matches exactly w, producing S_x. Extract all synsets $s_{m,j}$ in S_m (for word m) whose first lemma $l_{j,1}$ matches exactly m, producing S_y. (First part of 2 up to this point.) (b) Expand lemmas from both sets: (i) collect all lemmas for all synsets in S_x, producing L_x; (ii) collect all lemmas for all synsets in S_y, producing L_y. (c) Extract synsets for these lemma expansions: (i) collect all synsets for all lemmas in L_x, producing S_a; (ii) collect all synsets for all lemmas in L_y, producing S_b. (d) Return all unique lemmas $l_{c,1}$ of any synset $s_{a,c}$ in S_a that matches a synset $s_{b,d}$ in S_b using wup_similarity w/ threshold 0.9 (if any). Example: concerned → concerned.a.01 → implicated → implicated.s.01

Moral 4: (Added 5623 terms) This lexicon expansion yields a set of terms drawn from **all** synsets and their descendents that are semantically similar (above a threshold) to synsets and their descendents of words in the strong baseline (Moral 1) lexicon—without lemma matching—using the following steps: (a) Extract all synsets $s_{w,i}$ in S_w (for word w), producing S_x. Extract all $s_{m,j}$ in S_m (for word m), producing S_y. (b) Expand lemmas from both sets: (i) collect all lemmas for all synsets in S_x, producing L_x; (ii) collect all lemmas for all synsets in S_y, producing L_y. (c) Extract synsets for these lemma expansions: (i) collect all synsets for all lemmas in L_x, producing S_a; (ii) collect all synsets for all lemmas in L_y, producing S_b. (d) Return all unique lemmas $l_{c,1}$ of any synset $s_{a,c}$ in S_a that matches a synset $s_{b,d}$ in S_b using wup_similarity w/ threshold 0.9 (if any) Example: concerned → refer.v.02 → refer → pertain.v.02

B Moral Dimension/Value Lexicons: Moral 2-4

Three moral lexicons are induced automatically via propositional analysis, SRL, and semantic lexicon expansion:

- Moral 2: Moral 2 Lexicon
- Moral 3: Moral 3 Lexicon
- Moral 4: Moral 4 Lexicon

[5] https://github.com/oaraque/moral-foundations.

C Annotation Guidelines

1. For each tweet excerpt in the "Text" column A, apply steps 2–8 below.

2. Columns B through K are the potential Concern types. Put a "1" in a single column corresponding to the applicable Concern type. Leave empty if none appears to apply.

3. Columns L through W are moral dimensions. Put a "1" in any column that has an applicable moral dimension, in either the vice subcolumn or the virtue subcolumn. Leave the vice/virtue cells empty for any moral dimension that is not applicable. Refer to the moral dimensions described in the Graham and Haidt tables for this task (see links in 4a and 4b below).

4. Provide annotations only for explicitly represented material. Do not infer context that is not stated and do not apply any subject-matter background knowledge. The only background knowledge to be used for this task is the moral dimensions description in tables below from:

   (a) [Graham and Haidt, 2012] (see page 16)

   (b) [Graham et al., 2013] (see page 68)

5. Do not try to force the Text into a particular moral dimension. If a moral dimension appears to be applicable, but it is unclear whether the vice or virtue is active, put "DK" in either the vice cell or virtue cell, instead of leaving it blank.

6. Consider only the content of the Text without regard to grammaticality or punctuation.

7. Assume no sarcasm is present. Annotate the literal sense.

8. Use the last column (optionally) for any annotation notes, eg, the reasoning behind the chosen annotations.

D Ground Truth

We conduct an annotation task to develop Ground Truth (GT), against which to compare our concern detection system variants, based on 50 held-out tweets from the held-out development portion of the Kaggle English Twitter dataset on 2017 French elections (Daignan, 2017). Annotation was completed by two non-algorithm developers (one with expertise in linguistics, the other with expertise in psychosocial moral indicators). Annotators were provided guidelines (Appendix C) for both concern types and moral dimensions. For concern types the interrater reliability (IRR) is calculated through macroaveraging of kappa scores (Carletta, 1996) which produced a 66% agreement, considered **Strong** according to (McHugh, 2012). By contrast, the macroaveraged IRR for moral dimensions is considered none to slight (16%), which is deemed **Weak**. [Link to GT]

References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184.

Kathryn Baker, Michael Bloodgood, Bonnie Dorr, Chris Callison-Burch, Nathaniel Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Use of modality and negation in SIMT. *CL*, 38.

Albert Bandura, Claudio Barbaranelli, Gian Vittorio Caprara, and Concetta Pastorelli. 1996. Mechanisms of moral disengagement in the exercise of moral agency. *Journal of personality and social psychology*, 71(2):364.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Kate Crawford and Trevor Paglen. 2021. Excavating ai: The politics of images in machine learning training sets. *AI & SOCIETY*, pages 1–12.

Jean-Michel Daignan. 2017. French presidential election: Extract from twitter about the french election, kaggle data set.

Rosario Delgado and Xavier-Andoni Tibau. 2019. Why cohen's kappa should be avoided as performance measure in classification. *PLOS ONE*, 14(9):1–26.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press.

Scott E Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Christopher Miller, et al. 2021. Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7780–7788.

Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Online. Association for Computational Linguistics.

Jamie Lynn Goodwin, Andrew Lloyd Williams, and Patricia Snell Herzog. 2020. Cross-cultural values: a meta-analysis of major quantitative studies in the last decade (2010–2020). *Religions*, 11(8):396.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean Philip Wojcik, and Peter H. Ditto. 2012. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96 5:1029–46.

Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena P. Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101 2:366–85.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not weird. *Nature*, 466(7302):29–29.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength NLP in Python.

Ronald Inglehart and Chris Welzel. 2010. The wvs cultural map of the world. *World Values Survey*.

Ravi Iyer, Spassena Koleva, Jesse Graham, Peter Ditto, and Jonathan Haidt. 2012. Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PLOS ONE*, 7(8):1–23.

John T Jost, Jack Glaser, Arie W Kruglanski, and Frank J Sulloway. 2003. Political conservatism as motivated social cognition. *Psychological bulletin*, 129(3):339.

Peter Krejzl, Barbora Hourová, and Josef Steinberger. 2017. Stance detection in online discussions. *ArXiv*, abs/1701.00504.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Brodie Mather, Bonnie J. Dorr, Owen Rambow, and Tomek Strzalkowski. 2021. A general framework for domain-specialization of stance detection A covid-19 response use case. In *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624, Online. Association for Computational Linguistics.

Peter Pirolli, Kathleen Carley, Adam Dalton, Bonnie Dorr, Christian Lebiere, Michael Martin, Brodie Mather, Konstantinos Mitsopoulos, Mark Orr, and Tomek Strzalkowski. 2021. *Mining Online Social Media to Drive Psychologically Valid Agent Models of Regional Covid-19 Mask Wearing*, pages 46–56. 2021 International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SPB-BRIMS 2021).

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Younes Samih and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646, Online. Association for Computational Linguistics.

Shalom H. Schwartz. 2012. An overview of the schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2:11.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 579–589, Melbourne, Australia. Association for Computational Linguistics.

Asher Stern and Ido Dagan. 2014. Recognizing implied predicate-argument relationships in textual inference. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–744, Baltimore, Maryland. Association for Computational Linguistics.

Muhammad Umer, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. 2020. Fake news stance detection using deep learning architecture (cnn-lstm). *IEEE Access*, 8:156695–156706.

Jay J Van Bavel and Andrea Pereira. 2018. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, 22(3):213–224.

Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. 2020. Explainable deep learning: A field guide for the uninitiated. *ArXiv*, abs/2004.14545.

Brian Xu, Mitra Mohtarami, and James R. Glass. 2019. Adversarial domain adaptation for stance detection. *CoRR*, abs/1902.02401.

Bishan Yang and Claire Cardie. 2013. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2019. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5895–5906, Florence, Italy. Association for Computational Linguistics.