A General Framework for Domain-Specialization of Stance Detection: A Covid-19 Response Use Case

Brodie Mather and Bonnie J. Dorr
Florida Institute for Human and Machine Cognition, Ocala, FL, {bmather,bdorr}@ihmc.org
Owen Rambow
Stony Brook University New York, Stony Brook, NY, owen.rambow@stonybrook.edu
Tomek Strzalkowski
Rensselaer Polytechnic Institute, Troy, NY, tomek@rpi.edu

Abstract

We present a generalized framework for domain-specialized stance detection, focusing on Covid-19 as a use case. We define a *stance* as a *predicate-argument* structure (combination of an action and its participants) in a simplified one-argument format, e.g., *wear(a mask)*, coupled with a task-specific belief category representing the *purpose* (e.g., *protection*) of an argument (e.g., *mask*) in the context of its predicate (e.g., *wear*), as constrained by the domain (e.g., Covid-19). A belief category PROTECT captures a belief such as "masks provide protection," whereas RESTRICT captures a belief such as "mask mandates limit freedom." A stance combines a belief proposition, e.g., PROTECT(*wear(a mask)*), with a sentiment toward this proposition. From this, an overall positive attitude toward mask wearing is extracted. The notions *purpose* and *function* serve as natural constraints on the choice of belief categories during resource building which, in turn, constrains stance detection. We demonstrate that linguistic constraints (e.g., *light verb* processing) further refine the choice of predicate-argument pairings for belief and sentiment assignments, yielding significant increases in F1 score for stance detection over a strong baseline.

1 Introduction

Categorization of predicate-argument pairs (combinations of actions and their participants) from text input is a long-standing information-extraction task (Sundheim 1995; Chen et al. 2015; Liu, Luo, and Huang 2018) with recent advances leveraging contextualized representations (Du and Cardie 2020). The related task of event detection applies categorization to predicate-argument pairs for determining semantic equivalence and coreference (Song et al. 2016), as in *they wore masks* and *masks covered their faces*. However, many complex and fundamental challenges remain.

One challenge for such tasks is the need for a high degree of manual labor for domain tuning, leading to a limited set of categories overfit to a small domain. This issue is well recognized in the event detection community, e.g., in TAC-KBP's event nugget evaluation (Mitamura, Liu, and Hovy 2017). As another example, in the task of *ask detection* (Bhatia et al. 2020; Dorr et al. 2020), predicate-argument structures extracted from a social engineer's input are assigned categories such as PERFORM (if a social engineer attempts to get a potential victim to click a link, as in *To win $500 click this link*) or GIVE (if a social engineer attempts to get a potential victim to provide account information, as in *Enter your account number to win a free concert ticket*). Such approaches fall short in that there is no generalized framework for rapid ramp-up to new domains.

Another challenge is that underlying beliefs and sentiment (or attitudes) are generally not considered central to such tasks. However, these are critical for determining overall attitude not just toward entities in the world (e.g., *masks*) but also toward *functions* associated with entities (e.g., *a mask is worn*) that hinge on beliefs about *purpose* (e.g., *a mask protects*). Recent improvements in natural language processing (NLP) techniques, including more robust parsing and semantic role labeling (SRL) (Honnibal et al. 2020; Gardner et al. 2017), provide a basis for enriched predicate-argument extraction that supports induction of belief/sentiment-based attitudes toward topics of interest in a given domain.

We present a generalized framework for domain-specialized stance detection, focusing on Covid-19 as a use case. The approach involves categorization of predicate-argument pairs, with general linguistic constraints for increased accuracy. We define *stance detection* as a combination of belief, sentiment, and attitude in tweets about domain-relevant topics, e.g., mask wearing. A *stance* consists of a *predicate-argument* structure in a simplified one-argument format, e.g., *wear(a mask)*, coupled with a task-specific belief category representing the *purpose* (e.g., *protection*) of an argument (e.g., *mask*) in the context of its predicate (e.g., *wear*), as constrained by the domain (e.g., Covid-19). A category PROTECT captures a belief such as "masks provide protection," whereas RESTRICT captures a belief such as "mask mandates limit freedom." A stance combines a belief proposition, PROTECT(*wear*(*a mask*)), with a sentiment toward this proposition. The *stance* for *Wear a mask!* has a high belief strength (+3), and positive sentiment (+1): <PROTECT(wear (a mask)),+3,+1>. From this an overall positive attitude toward mask wearing is computed as the product of belief strength and sentiment: +3.

We show that the notions of *purpose* and *function* provide natural constraints on the choice of task-specific belief categories during resource building which, in turn, con-

strains stance detection. We demonstrate that linguistic constraints (e.g., *light verb* processing) further refine predicate-argument pairings and belief/sentiment assignments, yielding significant F1 score increases over a strong baseline.

Next, we present background and related work that inform our research. We then present our resource building methodology and stance detection, followed by our experimental design for both, and then results and discussion.

2 Background and Related Work

Producing stances that are indicative of individuals' beliefs is a central contribution of this work. Stance detection ingests various input sources (e.g., tweets, emails, and other textual based messages) and builds stances (and corresponding attitudes) from belief, belief strength, and sentiment. Although stance detection applied to social media is not new (AlDayel and Magdy 2020), we adopt a richer notion of "stance" here, beyond simple opinion mining.

Specifically, we apply dependency parsing and SRL to produce a structure that includes beliefs, belief strength (ranging from -3.0 to +3.0), and sentiment (ranging from -1.0 to +1.0). The belief strength scale aligns with Factbank (Saurí and Pustejovsky 2009): certain (+3.0), probable (+2.0), possible (+1.0), uncertain (0.0), unlikely (-1.0), improbable (-2.0), impossible (-3.0). These values are drawn from a small domain-independent lexicon adapted from our prior work on committed belief (Prabhakaran, Ganeshkumar, and Rambow 2018), and modality and negation (Baker et al. 2012). Sentiment is derived through composition of lexical terms from a small sentiment lexicon of general positive/negative terms (e.g., *like*, *hate*; see (Levin 1993)), negation terms, and domain-specific terms such as *protect* (inherently positive) and *restrict* (inherently negative). For example, *I don't like wearing masks* yields a negative sentiment toward mask wearing.

Following Ajzen (1991), an attitude score is computed for a topic, e.g., *mask wearing*, as the product of the belief strength and the sentiment toward that belief, were it to be true. The interaction between belief strength and sentiment is illustrated in the following cases: (a) *Wearing a mask definitely protects me* yields a strongly positive belief strength (3.0) toward mask protectiveness, with positive sentiment (1.0) if protectiveness is true; (b) *Wearing a mask doesn't protect me* yields a strongly negative belief strength (-3.0) toward mask protectiveness, with positive sentiment (1.0) if protectiveness is true; and (c) *Wearing a mask restricts my freedom* yields a strongly positive belief strength (3.0) toward mask restrictiveness, with negative sentiment (-1.0) if restrictiveness is true. The attitudes toward mask wearing are thus 3.0, -3.0, and -3.0.

We focus on population responses to Covid-related interventions as a use case for demonstrating generalizability of domain-specialized stance detection. As Covid-19 continues to have drastic global effects, it has become increasingly important to derive a sense of how people feel regarding critical interventions such as *mask wearing* or *social distancing*, especially as trends in online activity may be viewed as proxies for the sociological impact of COVID-19 (Sanders et al. 2020). Taking the pulse of a given population on topics such as those related to Covid-19 may predict how the public will handle restrictive situations and what actions need to be taken/promoted in response to emerging attitudes.

We design and implement a framework for rapid ramp-up of domain specialized stance detection. Although Covid-19 response is the domain of interest, the framework is applicable to tasks in other domains, such as *ask detection* in the social engineering domain (Bhatia et al. 2020; Dorr et al. 2020) or potentially in new future tasks involving discovery of emerging trends or misinformation.

There are two operative principles in our generalized approach to domain-constrained resource building:

- **Contribution 1**: Centrality of *function* and *purpose* in selection of belief categories for resource building;
- **Contribution 2**: Application of *one sense per domain/content* in trigger-content pairings for stances.

Following generative lexicon theory (Pustejovsky 1995), we adopt the notion of *qualia* to incorporate roles associated with domain-specific objects/entities such as *masks*: (1) **formal**, characterizing a mask as an article of clothing; (2) **constitutive**, characterizing a mask as having material, multiple folds, tight ties/elastics; (3) **telic**, characterizing a mask as having a potential *function* of being worn (over nose and mouth) and *purpose* of protection against particles (e.g., viral, dust); and (4) **agentive**, characterizing the creation of a mask via a factory or an individual. Of these, **telic** is viewed as central to resource building for stance detection, in that *function* and *purpose* underlie the belief categories associated with stances, e.g., PROTECT or RESTRICT. Thus, a term such as *wear* is tied to a belief category in the domain-constrained resource and stance detection then considers these terms to be potential indicators of such beliefs.

Another operative principle is a new notion of "One Sense per Domain/Content," an adaptation of "One Sense per Discourse/Collocation," (Yarowsky 1995) for accurate determination of *meaning* and extraction of stances. That is, the *sense* of a predicate (e.g., *wear, don, put on*) is tied to the argument's *function* with respect to its argument (*mask*), thus licensing an agent's belief about the argument's *purpose*, i.e., PROTECT or RESTRICT. A small set of such belief categories—motivated by the telic role of the argument (e.g., a *mask* is protective)—is associated with pairs of *triggers* (i.e., potential predicates such as *wear*) and *content* (i.e., potential arguments such as *mask*) by a human, during resource construction. These categories are posited based on the constrained nature of the domain (Covid-19 in our use case). That is, the telic role is deemed an appropriate (inferred) belief, as long as the appropriate domain-relevant context is available, e.g., *wear* is tied to PROTECT when paired with its content (*mask*) in the Covid domain.[1]

3 Domain-Specialized Resource Building

Although our domain focus is Covid-19, our resource-building methodology is general enough to be applied to any domain with naturally occurring data. To build a lexical resource for stance detection, we leverage the IEEE geo-

[1]By contrast, *Wear a costume* would not be assigned a belief category for stance detection in this domain.

tagged coronavirus Twitter data set (Lamsal 2020) (16,729 out of 300,000+ tweets), using 2450 held-out tweets for resource building and a held-out set of 50 tweets for our evaluation. These data contain a daily updated list of tweet ids that monitor real-time Twitter feeds for coronavirus-related geo-tagged tweets across the globe.

The key to generality for domain specialization is minimization of human labor (3.5 hours for Covid-19). Automatically suggested content words are reviewed by a human for assignment of a small set of belief categories, based on function/purpose (the telic role), e.g., PROTECT and RESTRICT for *mask*. Leveraging the "one sense per domain/content" constraint, trigger-content suggestions are presented to a human checker (a computational linguist who is not the stance-detection implementer) who confirms/rejects belief categories and assigns default belief strengths and sentiments (e.g., PROTECT defaults to +3 and positive sentiment +1). These same categories are automatically inherited by corresponding trigger words that represent the associated domain-specific **function** (e.g., *wear* in the case of *mask*). Belief categories are shown here with representative content terms and triggers (in parentheses):

- PROTECT: cloth, covering, sanitizer, mask (wear, sanitize, improvise, don)
- RESTRICT: distancing, lockdown, policy, mask (enforce, impose, mandate)
- SPREAD_ILLNESS: covid, coronavirus, C19, virus (contract, infect, spread)
- APPLY_MED_SCI: cure, infection, vaccine (measure, administer)
- ADAPT_LIFE: art, camping, hairstylist (advertise, bear, network)
- PROMOTE: biden, drfauci, trump (cheer, mock)
- BUILD_SOC: community, worldwide (divide, support, understand)
- TEST_TRACE: contact, tracer, contacttracing (trace, test)

Automation of content suggestions starts with parsing and SRL on 2450 held-out domain-relevant tweets, yielding 3450 potential content terms. A pre-processing step removes 2702 (78%) of these terms: (a) de-duplification of repeated terms; (b) removal of special characters and URLs; (c) removal of terms identified as sentiment or modality (*need* or *like*) that are handled independently via modality/sentiment processing; (d) removal of common closed-class words (*your, with*); (e) elimination of uncommonly appearing terms (*vocals*, *vending*) below threshold, e.g., 1-2 occurrences per 2450 tweets; and (f) elimination of words that appear most commonly as triggers (*visit*).[2]

The remaining 748 words that appear in argument positions of SRL output are suggested as content candidates to the human, who assigns a small set of domain-specific belief categories representing the *purpose* of relevant content words. Because of the constrained nature of the domain, human inspection of these words takes 2 hours: (1) 8 belief categories are posited for 50 representative Covid-19 content words; (2) the remaining 698 words are easily associated with these belief categories based on their similarity with other categorized content words (*coronavirus* and *covid*).

We focus here on PROTECT and RESTRICT (for *mask wearing*). The other 6 belief categories above are analogously assigned to content words, e.g., *virus* has *spread of illness* as its purpose and *vaccine* has *application of science/medicine* as its purpose. Content words may be associated with more than one belief category during resource building, accommodating multiple perspectives. In such cases, the trigger word is used to decide between them during stance detection: *wear a mask* is associated with *protection*, whereas *mandate masks* is associated with *restriction*.

An additional process yields a similar belief categorization for trigger words, but in this case the operative role is *function*, e.g., the trigger *wear* has a functional role with respect to its associated content (i.e., a *mask* is *worn*). First, predicates are automatically extracted from the parsed/SRL-processed tweets that contain belief-categorized content terms. The resulting 525 candidate triggers are automatically assigned belief categories of the associated content terms. Of these, 17 modality/sentiment terms (e.g., *desire*) are removed, as these are independently handled. The remaining 508 potential triggers are pared down to 268 via 1.5 hours of human inspection and elimination of the following cases: (a) incorrect lexical item chosen as the predicate by SRL for a given belief category, e.g., *add* is assigned a PROTECT trigger in *I'd like to add wear a mask to this sign*; (b) misanalyzed lexical items, e.g., *antiques* is considered a verb and thus is assigned a PROTECT trigger in a tweet about masks. The human verifier also specifies positive and negative valence to the 268 potential triggers. For example, *free* is a trigger word that implies a negative RESTRICT belief and *combat* implies a negative SPREAD_ILLNESS belief.

4 Stance Detection

Stance-detection extracts stances using both domain-specific trigger/content resources and domain-independent processing. The latter includes modality, negation, sentiment processing, and light verb processing, triggered by words such as *probably*, *definitely*, *not*, *like*, *hate*, and *use*. These processes are independent from, but superimposed on, beliefs and attitudes. The interplay between belief/sentiment and domain-independent linguistic constraints is shown here for positive and negative attitudes toward mask wearing:

Positive attitudes toward mask wearing
• *I wear/wore a mask*: <PROTECT(wear (masks)),+3,+1>
• *I like wearing masks*: <PROTECT(wear (masks)),+2.5,+1>
• *Wearing a mask definitely protects me*: <PROTECT(wear (masks)),+3,+1>
• *Wearing a mask probably protects me*: <PROTECT(wear (masks)),+2.5,+1>
• *I use a mask to stay safe*: <PROTECT(use (mask)),+3,+1>
• *I like masks*: <EXIST(masks),+3,+1>
Negative attitudes toward mask wearing
• *I don't/didn't wear masks*: <PROTECT(wear (masks)),-3,+1>
• *I don't like wearing masks*: <PROTECT(wear (masks)),+2.5,-1>
• *Wearing a mask bothers me*: <PROTECT(wear (masks)),+2.5,-1>
• *Masks restrict freedom*: <RESTRICT(restrict(mask freedom)),+3.0,-1>
• *I hate masks*: <EXIST(masks),+3,-1>

For example, the stance for *Wearing a mask definitely protects me* includes a PROTECT belief category triggered by *wear*, with *mask* as the corresponding content entry. The resulting stance includes a high degree of committed belief (+3), and a *positive* sentiment (+1), yielding an overall positive attitude toward mask wearing ($+3 \times +1 = +3$): <PROTECT(*wear(a mask)*), +3, +1>. However, if the term *probably* is used, a separate level of domain-independent

[2] 19 overlapping cases of trigger and content are retained: cover, mask, protect, sanitize, save, sew, shop, wash, wear, curb, fight, spread, close, distance, mandate, flatten, measure, online, neighbor.

modality processing averages the committed belief (+3) with modal belief for *probably* (+2), yielding a belief value of +2.5 for *Wearing a mask probably protects me*. Negative attitude is derived either from the belief value (e.g., in the first negative example, not wearing masks may be indicative of a belief that masks are not protective, i.e., -3) or from the sentiment (e.g., in the second negative example, the negation word *don't* associated with sentiment *like* yields -1, and *like* reduces the belief strength to +2.5 without flipping polarity).

Selection of PROTECT above follows from two operative principles: (1) The centrality of *function* and *purpose* drives the choice of such belief categories for domain-specialized resource building, i.e., PROTECT is deemed an appropriate belief type for the wear-mask pair because *wear* is a mask *function*, and masks serve the *purpose* of protection in the Covid-19 domain; (2) The centrality of one sense per domain/content licenses the selection of PROTECT as the belief category in the stance output, given that the trigger and content are constrained to bear the same belief category. Adherence to these principles during resource building and stance detection amounts to a form of mutual constraint whereby the trigger and target jointly and unambiguously constrain the domain-specialized belief category. This same constraint analogously enables the selection of RESTRICT for *Wearing a mask restricts my freedom.*

Another constraint incorporated into stance detection is linguistic in nature, i.e., *light verb* handling, which is an independent process that further enhances domain-specialized belief category selection. In *light verb* constructions, a semantically "light" predicate is coupled with a domain-specific argument that conveys the core meaning. Stance detection accommodates 9 light verbs, *be, do, give, have, make, put, take, use, place*. A sentence such as *I use a mask to stay safe* is interpreted by ignoring the light verb predicate (*use*) and selecting a belief based on the argument—in this case *mask*, which is associated with PROTECT during resource building. Light verb handling results in a significant reduction in false negatives in our experimental results.

Stance detection also includes a variant of predicate-argument processing similar to light verb processing, where sentiment is expressed without a belief trigger. For example, *I hate masks* does not convey belief about function or purpose, as would be the case for PROTECT or RESTRICT. Following (Russell 1919), the belief here is about *existence*, roughly referring to unhappiness that masks *exist*. The belief category is thus EXIST, with a belief strength of +3. The sentiment value is lexically determined, in this case -1 for the word *hate*. Negation is also taken into account for such cases as well as straightforward cases such as *I don't like wearing masks*, which is assigned a negative sentiment.

5 Experimental Conditions and Metrics

Our experiments involve an intrinsic evaluation that compares the performance of three stance detection implementations to a strong baseline, as well as to each other. The baseline (Base) refers to a trigger-target approach without robust modality and sentiment processing: a belief, belief strength, and default sentiment are assigned to a span of text if an identified trigger or content matches a predefined belief category (e.g., PROTECT). Unconstrained filtering (UCF) incorporates robust modality and sentiment processing, but ignores the mutually constraining nature of the trigger-target pair, thus overgenerating stance outputs. For example, a trigger may be associated with a belief category (say, PROTECT) that is incompatible with the content belief category (say, RESTRICT or no belief category at all). Mutually constrained filtering (MCF) produces a stance only if the trigger and content are associated with the same belief category, thus reducing overgenerated stances, but at the expense of undetected viable stances. Lightly loosened filtering (LLF) recovers many undetected viable stances by accommodating light verbs and introducing the EXIST category. This variant increases the accuracy of stance output with only a slight increase in overgenerated stances.

- **Baseline** (Base): Assign belief and default sentiment based on match against *either* trigger *or* content, without modality or sentiment processing
- **Unconstrained filtering** (UCF): Assign belief/sentiment based on match against *either* trigger *or* content, taking into account modality/sentiment
- **Mutually Constrained filtering** (MCF): Assign belief/sentiment based on match against *both* trigger *and* content, taking into account modality/sentiment
- **Lightly loosened filtering** (LLF): Assign belief and sentiment based on mutually constrained filtering plus light verb accommodation

Comparisons among these variants require a Ground Truth (GT) against which the output of stance detection is evaluated. Construction of an adjudicated GT involves an automatic process followed by validation of both potential and actual stance outputs by a human (a computational linguist who is not the stance-detection implementer), using 50 held-out tweets from a Twitter data set (IEEE Data set (Lamsal 2020)). An adapted version of *pooled relevance* (Voorhees and Harman 2001) is applied to detected stances by running all four systems, ranging from high recall (e.g., the UCF, with 73 stance outputs) to high precision (e.g., MCF, with 35 stance outputs), as well as a more balanced system (e.g., LLF with 55 stance outputs). Of the 186 total possible predicate-argument pairs (in the 50 tweets), 98 are associated with stances by at least one system[3] yielding a GT containing 55 "correct" stances. Crucially, 6 of these are **not** produced by any of the four systems (e.g., incorrect belief category); however, the human does not need to search for these missing cases (false negatives), as they are automatically flagged as potential stances by virtue of the (incorrectly assigned) predicate-argument pair appearing in system output. Once the human deems these stance-worthy, the corrected stances are added to the GT.

It is critical that the human-adjudicated GT take into account not just the extraction of "stance-worthy" tweets, but also the assignment of belief category, belief strength, and sentiment during stance detection (SD).[4] True Positives (TP), False Negatives (FN), and False Positives (FP) for belief categories, belief strength, and sentiment are assigned as follows: A TP corresponds to SD output for a

[3]There could be as many as 4 different belief categories, belief strengths, and sentiments, per stance (one per system). However, not all systems produce all stances, and 74 stances are redundant across systems. With duplicate removal, only 136 adjudications are required to produce the 50-tweet GT for 4 system outputs.

[4]SD = Base or UCF or MCF or LLF, defined above.

"stance-worthy" span of text assigned a belief category, belief strength, **and** sentiment that match those of GT; A FN corresponds to SD output for a "stance-worthy" span of text assigned a belief category that does not match GT **or** whose belief strength and/or sentiment do not match GT (or both); A FP corresponds to SD output for a "non-stance-worthy" span of text that is (incorrectly) assigned a belief category. TP, FN, and FP are used to compute Precision, Recall, and F1 (Manning, Raghavan, and Schütze 2008):

P= $\frac{TP}{TP+FP}$, R= $\frac{TP}{TP+FN}$, F1= $\frac{2*P*R}{P+R}$

6 Results and Discussion

As is common in constraint-based solutions to many AI problems, an exact solution is not always possible, so the goal is to find an approximate solution that yields benefit within a tolerance range (Hulubei and Freuder 1999), i.e., an instance of the *The Goldilocks Problem*. For our purposes, the goal is to achieve precision (P) and recall (R) levels that are "just right," by decreasing false positives (FPs) and false negatives (FNs) as much as possible without a significant hit to true positives (TPs). We have achieved this outcome, having run all four stance detection variants over the 50 tweets from the IEEE data set. Results are shown in Table 1.

Version	TP	FP	FN	P	R	F
Base	20	42	35	32.26	36.36	34.19
UCF	40	33	15	54.79	72.73	62.50
MCF	32	3	23	**91.43**	58.18	71.11
LLF	43	12	12	78.18	**78.18**	**78.18**

Table 1: Intrinsic evaluation results of stance detection

Processing without modality/sentiment robustness yields a low F1 of 34.19 (Base). With modality/sentiment robustness added, unconstrained predicate-argument filtering (UCF) significantly shifts TPs and FNs, improving both precision and recall (an 82% increase in F1) but still retaining a large number of FPs. That is, UCF **over**generates. Mutually constrained filtering (MCF) of predicate-argument pairs yields a drastic reduction in FPs (a 67% improvement in precision), with a minimal drop in TPs. That is, MCF **under**generates. The "just right" condition emerges with light verb accommodation (LLF), which reduces FPs relative to Base and UCF, yielding a significant gain in TPs involving light verbs, but with an increase in FPs. Overall, MCF has the highest precision (91.43) due to the highly accurate predicate-argument constraint, but LLF has a very large gain in recall (78.18) due to inclusion of light verbs missed by MCF. As such, LLF's F1 of 78.18 is the highest among all runs (10% increase over MCF, 128% increase over Base).

We tested the significance of performance improvements and observed statistically significant error reductions at well below the 0.01 probability level (McNemar 1947) for three system pairings: (Base,UCF), (UCF,MCF), (UCF,LLF).[5] The reduction in total error rate of LLF as compared to MCF was small and not statistically significant because the changes that improved recall also reduced precision. However LLF achieves the best balance between precision and recall, as indicated by its superior F1 score.

Additional analysis reveals that Base incorrectly assigns <RESTRICT(stay (safe)),+3,-1> to *Stay safe and drink plenty of fluids*, but UCF, MCF, LLF leverage robust modality to assign the correct output: <PROTECT(stay (safe)),+3,+1>. UCF overgenerates content words, producing <PROTECT(wear(mask covering transport)),+3,+1> for *You must protect others by wearing a face mask covering on public transport*. MCF reduces content to the relevant term *mask*: <PROTECT(wear(mask)),+3,+1>. LLF produces <PROTECT(use(sanitizer)),+3,+1> for *Use hand sanitizer*, whereas light verbs are not handled by any of the other three variants.

The predicate-argument pairs leveraged for stance detection provide a foundation not just for expression of sentiment, as is the focus of prior stance-detection techniques (e.g., *agree*, *disagree* (Riedel et al. 2017)), but for expression of belief and belief strength, which yields more accurate results over state-of-the-art (SoA) sentiment detection (Gardner et al. 2017).[6]

The study above yields a large majority of positive attitudes toward mask wearing. To explore whether this is true of other data sets, we consider a social media data set of 20K relevant tweets from February to November 2020 provided by CMU CASOS, separated into four sections corresponding to four states, with (pos,neg) attitude values as follows: CA (3730,11984), FL (9495,3236), NY (8433,3653), and PA (9183,2853). We observe that positive attitudes are dominant, with the exception of CA where an anomaly arose: two negative tweets are disproportionately re-tweeted, adding 9771 to the total negative count (6818 for one tweet, and 2953 for the other). This anomaly notwithstanding, attitudes within these twitter data sets are overwhelmingly positive towards masks which may hint at a bias inherent to Twitter itself–an issue that needs to be addressed in selection of data sets for testing stance detection.

7 Conclusions & Future Work

We have presented a general framework for domain-specialization of stance detection and demonstrated its usefulness for a Covid-19 use case. Mutual constraints applied to stance detection yield a significant performance boost. Additional linguistic constraints provide further improvements (Table 1). Moreover, the methodology described herein functions as a tool for rapid domain adaptation, through general techniques that enable resource construction with minimal human labor.

Future work will apply this methodology to other tasks involving predicate-argument pairs, such as extraction of emerging trends or discovery of disinformation. Linguistically motivated enrichments include: (1) Exploration of other content roles beyond telic for their impact on belief

[5]Tested values were correct responses (TP or TN) vs. incorrect responses (FP or FN), to determine the significance of change in total error rate.

[6]SoA Sentiment analysis assigns a label at sentence level, and thus is more prone to error. Predicate-argument pairs enable multiple outputs per sentence and yield more accurate attitude assignment. See comparison: https://github.com/ihmc/FLAIRS34/blob/main/SoA%20Comparison.pdf.

categorization, e.g., formal, constitutive, agentive; (2) Exploration of language independence, e.g., porting of modality, sentiment, and light verbs for multilingual stance detection; (3) Exploration of richer belief structures such as nested belief (e.g., *I wear a mask to slow the spread of covid*) and identification of the holder of the belief; and (4) refinement of the automation process of selecting candidate trigger and content words for resource building, exploiting mutual constraints between trigger words and content words.

Another future exploration is that of correlating stance detection output with attitudes provided in marked-up data sets. The CASOS data set categorizes twitter data in terms of pro/con/mixed/neutral toward mask wearing. A possible next step is to correlate cases associated with positive/negative attitudes toward mask wearing with CASOS' pro/con categorizations.

Finally, additional intrinsic and extrinsic evaluations are needed: (1) validation of belief and sentiment assignments via Amazon Mechanical Turk; and (2) explorations of correlations between stance-computed attitudes and behaviors of a population, community, or individual, from mask wearing survey data (Reinhart 2020).

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2020-20092500001 and in part by Department of Defense (DoD), Applied Research Laboratory for Intelligence and Security (ARLIS) via 93208-29524201. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, ARLIS, DoD or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Ajzen, I. 1991. The theory of planned behavior. *Organizational behavior & human decision processes* 50(2):179–211.

AlDayel, A., and Magdy, W. 2020. Stance detection on social media: State of the art and trends.

Baker, K.; Bloodgood, M.; Dorr, B.; Callison-Burch, C.; Filardo, N.; Piatko, C.; Levin, L.; and Miller, S. 2012. Use of modality and negation in SIMT. *CL* 38.

Bhatia, A.; Dalton, A.; Mather, B.; Santhanam, S.; Shaikh, S.; Zemel, A.; Strzalkowski, T.; and Dorr, B. J. 2020. Adaptation of a lexical organization for social engineering detection and response generation. In *Proc. LREC STOC Workshop*, 9–14.

Chen, Y.; Xu, L.; Liu, K.; Zeng, D.; and Zhao, J. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of ACL/IJCNLP*, 167–176.

Dorr, B. J.; Bhatia, A.; Dalton, A.; Mather, B.; Hebenstreit, B.; Santhanam, S.; Cheng, Z.; Shaikh, S.; Zemel, A.; and Strzalkowski, T. 2020. Detecting asks in social engineering attacks: Impact of linguistic/structural knowledge. In *AAAI*, 7675–7682.

Du, X., and Cardie, C. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of EMNLP*, 671–683. Online: Association for Computational Linguistics.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. S. 2017. Allennlp: A deep semantic NLP platform.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength NLP in Python.

Hulubei, T., and Freuder, E. C. 1999. The goldilocks problem. In Jaffar, J., ed., *Principles and Practice of Constraint Programming – CP'99*, 234–245. Springer Berlin Heidelberg.

Lamsal, R. 2020. Coronavirus (covid-19) geo-tagged tweets.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Liu, X.; Luo, Z.; and Huang, H. 2018. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1247–1256. Brussels, Belgium: Association for Computational Linguistics.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. USA: Cambridge U. Press.

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

Mitamura, T.; Liu, Z.; and Hovy, E. H. 2017. Events detection, coreference and sequencing: What's next? overview of the TAC KBP 2017 event track. In *Proceedings of the Text Analysis Conference*. NIST.

Prabhakaran, V.; Ganeshkumar, P.; and Rambow, O. 2018. Author commitment and social power: Automatic belief tagging to infer the social context of interactions. In *Proceedings of NAACL-HLT (Long Papers)*, 1057–1068.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Reinhart, A. 2020. New and improved covid symptom survey tracks testing and mask-wearing.

Riedel, B.; Augenstein, I.; Spithourakis, G. P.; and Riedel, S. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR* abs/1707.03264.

Russell, B. 1919. General propositions and existence: The philosophy of logical atomism, lect 5. *The Monist 29* 190–206.

Sanders, A. C.; White, R. C.; Severson, L. S.; Ma, R.; McQueen, R.; Paulo, H. C. A.; Zhang, Y.; Erickson, J. S.; and Bennett, K. P. 2020. Unmasking the conversation on masks: NLP for topical sentiment analysis of covid-19 twitter discourse.

Saurí, R., and Pustejovsky, J. 2009. Factbank: a corpus annotated with event factuality. *LREC* 43:227–268.

Song, Z.; Bies, A.; Strassel, S.; Ellis, J.; Mitamura, T.; Dang, H.; Yamakawa, Y.; and Holm, S. 2016. Event nugget and event coreference annotation. 37–45.

Sundheim, B. 1995. Overview of results of the muc-6 evaluation. volume 1996, 13–31.

Voorhees, E., and Harman, D. 2001. Overview trec-9. *Information Processing and Management* 31.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of ACL*, 189–196.