Toward Comprehensive Understanding of a Sentiment Based on Human Motives

Naoki Otani
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
notani@cs.cmu.edu

Eduard Hovy
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
hovy@cmu.edu

Abstract

In sentiment detection, the natural language processing community has focused on determining holders, facets, and valences, but has paid little attention to the *reasons* for sentiment decisions. Our work considers human motives as the driver for human sentiments and addresses the problem of motive detection as the first step. Following a study in psychology, we define six basic motives that cover a wide range of topics appearing in review texts, annotate 1,600 texts in restaurant and laptop domains with the motives, and report the performance of baseline methods on this new dataset. We also show that cross-domain transfer learning boosts detection performance, which indicates that these universal motives exist across different domains.

1 Introduction

Understanding a person's sentiment based on text has practical implications for improving product/service quality, along with scientific implications for psychology and other fields. Despite a rich body of sentiment analysis research, a sentiment is often simply assumed to be expressed by uni-dimensional binary or ternary labels (positive, neutral, and negative), and relatively little attention has been paid to the reason for holding a particular sentiment value. Aspect-based sentiment analysis (ABSA), which considers fine-grained categories (a.k.a. aspects) that may cause sentiment, partially tackles this problem. However, aspects are typically limited to properties of entities such as the price of food and design of a product (e.g., (Pontiki et al., 2016)) and do not really show *why* such aspects matter and *how* they cause human sentiments. For example, some people desire cheap and quick meals for saving time and money, and others desire high-grade food for enjoying the dining experience itself.

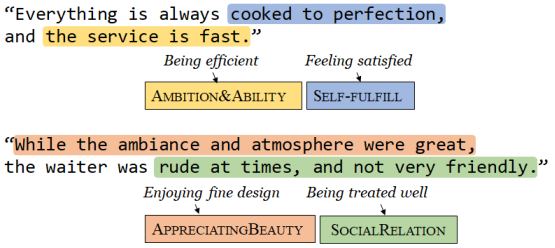

Figure 1: Restaurant review texts and human motives of interest (rectangles).

Following Li and Hovy (2017), we consider a sentiment as a realization of an individual's mental state that relates to his/her satisfaction toward a specific event or entity. While a sentiment can be driven by a sentiment holder's emotional, non-logical preference (like "I just don't enjoy that kind of food") and also conditioned by long-term plans and resources that the holder has, a sentiment is largely triggered by whether one of the holder's goals is satisfied or not. As Figure 1 illustrates, one will have a negative sentiment toward a restaurant if the service is terrible because one's basic motive for social behavior is not met.

What and how many motives do we have? Decades of effort have been devoted to this question in research areas such as psychology, for example, (Maslow, 1943). A recent study by Talevich et al. (2017) defines a taxonomy of motives, including SELF-FULFILLMENT, APPRECIATING BEAUTY, SOCIAL RELATION, HEALTH, AMBITION&ABILITY, and FINANCE. We use their comprehensive taxonomy for understanding sentiments.

Our work is in line with studies attempting to identify relevant motives in texts (Ding and Riloff, 2018; Rashkin et al., 2018), aiming to equip machines with the ability to understand a more complete description of a situation and justify human decisions and actions. While Ding and Riloff (2018) and Rashkin et al. (2018) specifically focus

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4672–4677
Florence, Italy, July 28 - August 2, 2019.

on predicate-argument tuples and artificial texts, respectively, our work analyzes real review sentences.

As an initial step, we conduct a task of human motive detection. We manually annotate 1,600 review texts in restaurant and laptop domains from existing ABSA datasets with the six motives. The annotation results reveal that people are driven by different motives in different domains. Finally, we report the performance of baseline methods on this new dataset. The results indicate a substantial space to improve automatic detection methods.

Following research on human motivation, we hypothesize that underlying drivers of human behavior are universal across domains, though distributions can vary. With this assumption, we leverage out-of-domain data to improve a human motive detector in the target domain. Our experiment indeed shows that transfer learning across restaurant and laptop domains is effective in motive detection.

2 Representation of Human Motives

Our aim is to justify a sentiment using human motives. To this end, we require a taxonomy of human motives. Motives are defined as reasons people hold for initiating and performing voluntary behavior (Reiss, 2004). A study of human motives dates back to Aristotle (384-322BC), who proposed a distinction between ends and means.[1] Ends, for which there are several theories, are believed to be a closed class (e.g. (Maslow, 1943)).

The aforementioned motives are drawn from a taxonomy of 161 motives (Talevich et al., 2017). Talevich et al. derived basic motives based on an extensive literature survey and grouped them hierarchically based on similarity judgments collected from human subjects. The hierarchical structure of their taxonomy embodies conceptual relationships between motives. Higher-level motives in the hierarchy are more abstract. The motives we picked are intermediate categories in the taxonomy that cover a wide range of topics appearing in our review texts (Table 1). These intermediate categories represent 55% of the taxonomy.

3 Annotation of Human Motives

We use Amazon Mechanical Turk to annotate review texts. We assign three crowd annotators to each text and aggregate their responses to obtain the final results.

3.1 Setup

Data: We annotate restaurant and laptop review texts from the SemEval 2016 datasets (Pontiki et al., 2016). We extract sentences with fewer than 25 tokens,[2] and sample 800 sentences from each domain.

Quality Control: We first collect annotations on 200 sentences in each domain without any filtering of workers. We then evaluate the workers on the 400 sentences: one of the authors examine the responses and made the gold-standard label set, and we calculate the F1-score of each worker against the gold-standard. We only use the workers whose scores are ≥ 0.5 in the remaining annotation tasks.

3.2 Results

Annotation Agreement: Our crowd workers agreed moderately on annotations: Krippendorffs α was 0.48 and 0.59 in the restaurant and laptop domains, respectively. We found that SELF-FULFILLMENT and EMBRACE & EXPLORE LIFE are often hard to distinguish. We, therefore, collapsed these categories, and Krippendorffs α increased to 0.51 and 0.61. For reference, three graduate students studying language technology annotated 150 sentences in the restaurant domain. Their Krippendorffs α was 0.72 on the original annotation scheme and 0.74 on the collapsed scheme.

Analysis: We next aggregated crowd workers responses using MACE (Hovy et al., 2013), where a response was regarded as a binary value of a combination of a text and a human motive. We set the prior probability of a positive class to 1/6 (i.e., one text is likely to have one of the six motives). This prior fits the responses better than a uniform prior.

Table 2 shows the distributions of human motive labels. There is a clear difference between domains: the restaurant domain has a variety of motives relevant to hedonic motives (i.e. pleasure seeking) like SELF-FULFILLMENT (SF) and SOCIAL RELATION (SR), while the laptop domain tends to have utilitarian motives (i.e. practical needs) such as AMBITION&ABILITY (AA) and FINANCE (F).

[1] In his book *"Nicomachean Ethics"*

[2] We use Stanford CoreNLP v.3.9.2 (Manning et al., 2014) to tokenize sentences.

SELF-FULFILLMENT (SF)	Finding meaning in life. Feeling satisfied with one's life.	"Ess-A-Bagel is by far the best bagel in NY."
*EMBRACE &EXPLORE LIFE (EE)	Being entertained. Exploring a new thing.	"The wine list is extensive."
APPRECIATING BEAUTY (BA)	Enjoying fine design/natural beauty. Being creative.	"A beautifully designed dreamy restaurant."
SOCIAL RELATION (SR)	Being treated well by others. Belonging to a social group.	"Everyone was cheerfully cooperative."
HEALTH (H)	Being physically healthy.	"The fish was not fresh and the rice tasted old."
AMBITION&ABILITY (AA)	Being competent/knowledgeable. Keeping things in order. Being efficient.	"I've waited over one hour for food."
FINANCE (F)	Saving money Getting things worth the financial cost.	"The prices are high, but I felt it was worth it."

Table 1: Motive categories, definitions and examples sentences. *EMBRACE&EXPLORE LIFE is merged to SELF-FULFILLMENT (Section 3.2).

	SF	AB	SR	H	AA	F
Restaurant	348	79	137	31	95	109
Laptop	188	164	52	9	370	145

Table 2: Distribution of human motives.

4 Human Motive Detection

We propose the task of motive detection. This is a multi-label sentence classification task, where for a given sentence a system detects relevant human motives. One text can have multiple labels.

4.1 Baseline Models

4.1.1 SVM

We run a linear SVM classifier on bag of n-grams (BoNG) of sentences. We count 1-, 2-, and 3-grams of words in each sentence to construct a BoNG vector. To avoid overfitting to rare words, we discard n-grams that occur only once in a training set. We also apply TF-IDF scaling to BoNG vectors to emphasize topic words ($\text{BoNG}_{\text{tfidf}}$).

4.1.2 Multi-layer Perceptron (MLP)

We build an MLP classifier with one hidden layer on top of word embedding-based sentence representations. We compress a variable-sized sequence of word embeddings into a fixed-sized sentence embedding before feeding them into MLPs using three standard encoders below.

Simple word-embeddings model (SWEM): We calculate element-wise average and max-pooling of word embeddings in a sequence and concatenate them (Shen et al., 2018).

CNN: A CNN aggregates adjacent word units in a hierarchical manner. We follow Kim (2014) and use filter windows of 3, 4, and 5.

Bidirectional LSTM (BiLSTM): A bidirectional LSTM encodes the whole word order in a sentence. We concatenate hidden states at the final time steps from both directions to obtain a sentence vector. We set the number of layers to two.

4.2 Training

We simply treat our multi-label classification task as a set of binary classification tasks, where MLP classifiers share parameters except for those of an output layer over motive categories. To handle highly skewed class distributions, we minimize a weighted loss function to train a model. For example, MLP classifier minimizes a weighted cross-entropy loss:

$$\mathcal{L} = -\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}}\sum_{c\in\mathcal{C}}\left[w_c y_c \log \text{MLP}_c(\boldsymbol{x}) + (1-y_c)\log(1-\text{MLP}_c(\boldsymbol{x}))\right], \quad (1)$$

where $(\boldsymbol{x}, \boldsymbol{y})$ is a pair of a sentence and a label in dataset $\mathcal{D}$, $\mathcal{C}$ is a set of categories, and MLP_c is an output function w.r.t. category c. We use the following class weight (Morik et al., 1999).

$$w_c = \frac{\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}}(1-y_c)}{\sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} y_c} \quad (c\in\mathcal{C}) \quad (2)$$

4.3 Transfer Learning Across Domains

In contrast to entity aspects that must be defined for each domain, underlying human motives will be universal across domains although distributions can be different. If this hypothesis is true, we can leverage out-of-domain data to improve motive detectors. We conduct transfer learning across

Method	Restaurant			Laptop		
	Precision	Recall	F1	Precision	Recall	F1
SVM-BoNG	.565 (±.028)	.394 (±.052)	.451 (±.046)	.480 (±.043)	.358 (±.018)	.397 (±.006)
+ Transfer	.518 (±.042)	.422 (±.055)	.459 (±.050)	.555 (±.102)	.396 (±.023)	.453 (±.039)
SVM-BoNG$_{\text{tfidf}}$	.544 (±.018)	.482 (±.032)	.492 (±.015)	**.577** (±.106)	.449 (±.014)	.477 (±.038)
+ Transfer	.542 (±.060)	.477 (±.041)	.475 (±.012)	.566 (±.071)	.461 (±.012)	.489 (±.029)
MLP-SWEM	.376 (±.027)	**.783** (±.002)	.478 (±.026)	.359 (±.007)	**.592** (±.022)	.416 (±.001)
+ Transfer	.462 (±.040)	.662 (±.025)	.516 (±.013)	.393 (±.011)	.534 (±.047)	.436 (±.025)
MLP-CNN	.565 (±.032)	.499 (±.045)	.524 (±.032)	.468 (±.011)	.410 (±.014)	.423 (±.007)
+ Transfer	**.700** (±.059)	.541 (±.017)	**.583** (±.020)	.519 (±.033)	.432 (±.021)	.456 (±.006)
MLP-LSTM	.447 (±.007)	.631 (±.008)	.511 (±.007)	.419 (±.007)	.568 (±.001)	.473 (±.005)
+ Transfer	.475 (±.017)	.618 (±.011)	.531 (±.005)	.500 (±.031)	.572 (±.003)	**.518** (±.006)
(Ref.) Human	.724 (±.014)	.859 (±.014)	.781 (±.012)	.766 (±.021)	.855 (±.019)	.806 (±.017)

Table 3: **Results of human motive detection.** Macro-precision, recall, and F1-measure scores are averaged over three folds in cross-validation (except for the performance of crowd workers in row Human). The higher numbers in each metric are denoted in **bold face**.

domains by minimizing the loss function below.

$$\mathcal{L}' = \mathcal{L}_{\text{in}} + \lambda \mathcal{L}_{\text{out}}, \tag{3}$$

where $\mathcal{L}_{\text{in}}$ and $\mathcal{L}_{\text{out}}$ are loss functions defined on in-domain and out-of-domain data, and λ is a hyperparameter to discount the out-of-domain loss.

5 Experiments

5.1 Setup

We use macro-averaging of F1 measures over motive categories as the primary evaluation metrics. We conduct three-fold cross-validation, where the dataset is divided evenly into training, validation, and test sets. In each fold, we conduct a grid search of hyperparameters based on the validation set. We then use a training and validation set to train a model and test on a test split. We report the average scores over test splits as the final score.

We use pretrained 100-D GloVe embeddings trained on 6 billion tokens from Wikipedia and Gigaword corpus (Pennington et al., 2014).[3] We provide the implementation details in the appendix.

5.2 Results

Table 3 shows that the MLP classifiers performed better or on par with the SVM classifier in terms of F1 measure. The low recall scores of SVM classifiers indicate that surface-level features are insufficient to detect various realizations of human motives.

[3] https://nlp.stanford.edu/projects/glove/

Interestingly, adding out-of-domain data improved F1 of all classifiers except SVM-BoNG$_{\text{tfidf}}$. Particularly, the precision of the MLP classifiers increased by transfer learning. For the MLP-CNN classifier, the boost from laptop domain instances was as high as 0.059 of F1 measure. This fact indicates the universality of underlying motives across domains.

We also report on human performance by comparing individual responses of crowd workers against aggregated, gold-standard labels. We generated 100 sets of human responses by repeatedly sampling one of three workers for each sentence. We can see a large gap between the classifiers (0.5 F1) and human (0.8 F1) in this task.

6 Discussion

We analyzed two domains, restaurants and laptops. In the restaurant domain, people are driven by hedonic motives in many cases and utilitarian motives in some cases. The laptop is a domain where people are driven by utilitarian motives in the majority cases. Of course, there are many domains other than these domains. For example, people would watch only for enjoying it. Exploring other domains would be an interesting direction for future research.

Another important direction is to develop a method that bridges between motives and sentiment valence. Li and Hovy (2017) gives concrete procedures to account for semantics behind the scene: we identify which goals are aimed at to fulfil a given motive, which plans are taken to

achieve the goals, which actions and conditions appear in the plans, and how well they are actually performed. These intermediate components would relate what we call *aspects* in aspect-based sentiment analysis.

Although we focused on sentiment analysis in this study, detection of motives can benefit other NLP applications such as in-depth machine reading. For example, underlying motives will be a strong clue for modeling a sequence of actions that share the same actor (a.k.a *narrative chains* (Chambers and Jurafsky, 2008)).

7 Conclusion

We aimed at understanding why a writer of a text holds a particular sentiment and proposed a task of human motive detection as an essential building block to this end. We presented a taxonomy of motives derived from a psychology study and annotated 1,600 restaurant and laptop reviews with six motives. We evaluated the performance of baseline predictive models on this dataset.[4]

One interesting property is that the same underlying motives can appear in different domains even though their distribution may differ. We empirically verified this by transferring learned parameters across domains. The result showed that predictive models can strongly benefit from out-of-domain instances. Nevertheless, there is still a substantial performance gap between humans and automatic detectors.

Acknowledgments

We thank Aldrian Obaja Muis and Hiroaki Hayashi for participating in our preliminary annotation exercise. We appreciate useful feedback from the anonymous reviewers.

References

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 789–797, Columbus, Ohio, USA. Association for Computational Linguistics.

Haibo Ding and Ellen Riloff. 2018. Human Needs Categorization of Affective Events Using Labeled and Unlabeled Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1919–1929, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1120–1130. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jiwei Li and Eduard Hovy. 2017. Reflections on Sentiment/Opinion Analysis. In *A Practical Guide to Sentiment Analysis*, pages 41–59. Springer, Cham.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 55–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Abraham H. Maslow. 1943. A Theory of Human Motivation. *Psychological Review*, 50(4):370–396.

Katharina Morik, Peter Brockhausen, and Thorsten Joachims. 1999. Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 268–277, Bled, Slovenia. Morgan Kaufmann Publishers.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

[4] Code and data used in this study are available online: https://github.com/notani/acl2019-human-motive-identification

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Steven Reiss. 2004. Multifaceted Nature of Intrinsic Motivation: The Theory of 16 Basic Desires. *Review of General Psychology*, 8(3):179–193.

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 440–450, Melbourne, Australia. Association for Computational Linguistics.

Jennifer R. Talevich, Stephen J. Read, David A. Walsh, Ravi Iyer, and Gurveen Chopra. 2017. Toward a Comprehensive Taxonomy of Human Motives. *PLOS ONE*, 12(2):e0172279.