SENPAI: Supporting Exploratory Text Analysis through Semantic & Syntactic Pattern Inspection

Mattia Samory
Computational Social Science
GESIS, Cologne, Germany
mattia.samory@gesis.org

Tanushree Mitra
Department of Computer Science
Virginia Tech, VA, USA
tmitra@vt.edu

Abstract

Analyzing language for social computing tasks requires looking beyond individual words. For example, the word "please" generally signals politeness, but more so together with modal verbs ("could you please...") than without ("please do this."). Combining semantics and syntax into rich textual patterns is essential to capturing these nuances. What are the relevant patterns for a task, and how to find them? NLP practitioners choose patterns informed by theory, and find them through computational models. However, few tools allow identifying rich patterns without NLP expertise. We introduce SENPAI, a novel tool that discovers combined semantic and syntactic patterns. SENPAI fuses neural embeddings, dependency parsing, and graph mining to surface patterns directly from data. We apply SENPAI to measure credibility, politeness, and sentiment in text. Quantitatively, models powered by SENPAI perform similarly to theoretically-motivated ones. Qualitatively, SENPAI discovers patterns that are interpretable and meaningful. SENPAI enables building computational models without NLP expertise and discovering new linguistic constructs.

1 A motivating example

What makes a good joke? Alice, a social computing scholar, wants to find out by studying submissions to *r/jokes*[1], a subreddit for humor. She has an intuition that not all jokes are completely original but instead follow certain patterns, such as "knock knock" and "lightbulb" jokes. Some patterns may make better jokes than others. However, which ones? Which patterns are even present in the submissions? To find out, Alice has several options. First, she can sample a few submissions and manually code them to surface patterns. This is a labor-intensive task and being a manual process, she might overlook important patterns. Second, she can review the linguistics literature on humor to learn what patterns jokes are supposed to use, translate patterns into code, and use a computational model to find the patterns in the submissions. This would allow Alice to analyze all submissions in r/jokes without the need to search manually. However, literature on humor is mostly informed by jokes outside of the social media context, e.g. stand-up comedy transcripts. Hence, a theory-driven pattern search would be constrained by what is present in the literature, while missing many patterns that might be otherwise common in the data; e.g., submissions in r/jokes that refer to internet meme jokes. Ultimately, Alice decides to use SENPAI (figure 1 summarizes how).

Alice feeds SENPAI the submissions from r/jokes. SENPAI automatically finds the patterns that appear frequently in the submissions. Alice looks at the patterns discovered by SENPAI, and she finds out that not only "knock knock" and "lightbulb" are common, but also "what do you call..." and "I don't always...but"[2]. SENPAI not only tells Alice what patterns are frequent in r/jokes, but also tells her, in fine detail, which patterns are present in each submission. This enables Alice to study which patterns make the best jokes in a particular social media context—Reddit's r/jokes community. She obtains ground truth on which jokes are the funniest by crowdsourcing annotations of hilariousness, on a scale from 0 (😒) to 5 (😂). She divides submissions into "funny" and "not funny" according to the annotations. Then, she trains a regression model on the average hilariousness score, using the patterns discovered by SENPAI as features. By inspecting the model, Alice can prove that the submissions using the "I don't always...but" pattern make the funniest jokes, whereas the "knock knock" pattern, popular in theory, is out of fashion in r/jokes. SENPAI allowed Alice to identify the patterns that characterize submissions in r/jokes and to incorporate them into a computational model of hilariousness, without the laborious process of manual coding or extensively browsing literature.

2 Introduction

Computational models of linguistic dimensions, such as credibility, politeness, and sentiment, manifest in the form of textual patterns (Cambria and White 2014; Ellis 2002). Patterns need to consider not only words in isolation, but also how words relate to one another. Combining semantics and syntax allows discerning, for example, the meaning of the word "mad" in the sentences "he is mad at you" and "he is a mad hatter." Which semantic and syntactic patterns are useful for measuring a linguistic dimension? Which of those patterns are relevant in a specific application domain?

Social computing researchers typically adopt one of two approaches to answer those questions. First, natural language processing (NLP) practitioners let theory of linguistics guide their choice of the most appropriate patterns for the task. They then

[1]examples taken from https://github.com/amoudgl/short-jokes-dataset

[2]https://knowyourmeme.com/memes/the-most-interesting-man-in-the-world

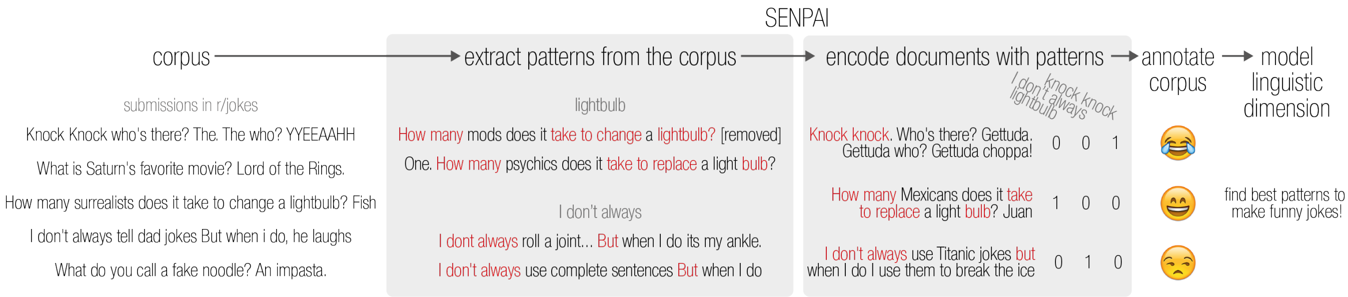

Figure 1: SENPAI discovers rich textual patterns that reflect linguistic dimensions such as hilariousness. Researchers can use SENPAI to analyze a corpus, e.g. a collection of jokes. SENPAI automatically discovers rich textual patterns that appear in the corpus. Since SENPAI finds combined semantic and syntactic patterns, it is able to detect words with similar meaning—e.g. "change" and "replace"—and patterns that span multiple words—e.g. "I don't always... but." SENPAI encodes each document in the corpus with the patterns it discovers. Thus, SENPAI enables not only exploring the linguistic properties of the corpus, but also building computational models of the linguistic dimensions, e.g. classifier for funny jokes, without NLP expertise.

leverage machine learning expertise to craft ad-hoc computational models. Yet, ad-hoc models suffer from selection bias: linguistics theory may not account for patterns that appear in the application domain. For example, social media text frequently include emojis (Barbieri et al. 2018) and regional variations (Eisenstein 2018)—or references to online memes in our motivating example. An alternative approach to ad-hoc models is to rely on general purpose tools, such as lexicons, word embeddings, and topic models. Social computing researchers rely on those tools to search for linguistic dimensions such as sentiment (Tausczik and Pennebaker 2010), humor (Kasunic and Kaufman 2018), and behavior change (Chancellor, Hu, and De Choudhury 2018). However, those general purpose tools are mostly limited to semantics, and miss on the deeper insights that come from combining them with syntax (Gildea and Palmer 2002). In sum, social computing researchers would benefit from tools with three properties: 1) provide insight beyond simple semantics, 2) are adaptable to the application domain, and 3) are easy to inspect and interpret.

With this paper, we provide just such a tool. SENPAI supports text analysis by facilitating **SE**mantic and sy**N**tactic **PA**ttern **I**nspection. SENPAI finds patterns automatically by leveraging state-of-the-art word vectors, neural embeddings, and statistical dependency parsing, thus unburdening the researcher from implementation details. Researchers can use SENPAI to gain insight on a linguistic dimension directly from data, such as sentiment in reviews. SENPAI finds patterns that best model that dimension, such as booster words ("the camera is *great*") and their relation to contrastive conjunctions ("the camera is *great, but* the battery is unusable"). On the one hand, SENPAI can serve as a black-box tool to discover patterns to use as features in computational models. On the other, researchers can use SENPAI to verify the presence of patterns suggested by theory and even discover domain-specific patterns.

After situating our work in related literature, we discuss how SENPAI discovers representative patterns automatically. We present three case studies: measuring credibility (Soni et al. 2014), politeness (Danescu-Niculescu-Mizil et al. 2013), and sentiment expressions (Hutto and Gilbert 2014) in online texts. The motivation behind the choice of case studies is threefold. First, they tackle central dimensions in sociolinguistics. Those dimensions are crucial in social computing because they underlie social processes such as signaling trustworthiness, hierarchy, and affect respectively. Second, those dimensions have been computationally modeled accurately, thus posing a strong baseline to beat. Third, the studies allow validating SENPAI on a variety of datasets, from tweets to long-form paragraphs. We validate that users without social computing or NLP expertise can make sense of the patterns discovered by SENPAI, through crowdsourcing. Human annotators can correctly interpret 87% of the patterns that SENPAI discovers automatically. We compare the patterns automatically discovered by SENPAI with the theoretically-motivated patterns that were manually compiled in the case studies. For all three linguistic dimensions, the patterns discovered by SENPAI include the theoretically-motivated patterns. Furthermore, through qualitative evaluations with social computing researchers, we show that SENPAI surfaces new patterns that are meaningful for studying the three linguistic dimensions. Ultimately, we show that substituting the patterns discovered by SENPAI for the theoretically-motivated patternsin the case studies results in more accurate computational models. We offer SENPAI as a system to enable faster, richer sociolinguistic research in the ICWSM community.

3 Related work

SENPAI draws on prior work on computational linguistics, and builds upon current semantic and syntactic analysis tools.

3.1 Measuring signals in text

Text can help understand human factors. Text reveals not only the identity of its author (Stamatatos 2009), but also nuanced characteristics like personality traits (Schwartz et al. 2013). Social media allowed investigating fundamental social processes through large-scale text corpora (Nguyen et al. 2016). For example, scholars gained an empirical understanding of linguistic signals of hierarchy (Gilbert 2012) and consensus (Baronchelli 2018). At an individual scale, psycholinguistics have looked, for example, at how deception (Donath 1999), persuasion (Tan et al. 2016), and memorability (Danescu-Niculescu-Mizil et al. 2012) affect everyday decisions.

When studying socio- and psycho-linguistic dimensions, scholars rely on theory to identify corresponding patterns in text, e.g. focusing on plural pronouns and assent to measure social coordination (Tausczik and Pennebaker 2010). However, identifying those patterns in text involves dealing with the sophisticated NLP techniques and specifics of the application domain (Nguyen et al. 2016). This is particularly taxing in exploratory studies, for example when platforms introduce new means of communication (e.g. quote-RT on Twitter (Garimella, Weber, and De Choudhury 2016)). Language online changes frequently (Nguyen, McGillivray, and Yasseri 2017; Pavalanathan and Eisenstein 2015) and application domain plays an important role in shaping it (e.g. hashtag variations to avoid moderation in Instagram (Stewart et al. 2017)). SENPAI facilitates exploratory studies by removing the need for complex computational processing and familiarity with conventions in the application domain.

3.2 Tools for computational linguistics

Lexicons like LIWC (Tausczik and Pennebaker 2010) and Empath (Fast, Chen, and Bernstein 2016) allow measuring the emotional (e.g. anger), grammatical (e.g. verb tense), and topical (e.g. religion) dimensions of text through simple word counts. One drawback of lexicons is that they ignore the context of the words they identify. For example, LIWC codes the word "mad" in the "anger" lexicon, and would therefore miscode it when it takes on other meanings, e.g. as a synonym of "crazy" in the phrase "mad scientist." More complex tools augment lexicons with syntactic information to disambiguate context (e.g. VADER (Hutto and Gilbert 2014)). These tools study specific linguistic dimensions and rely on ad-hoc syntactic patterns, for example differentiating whether the word "please" appears at the start or in the middle of a sentence as a signal of politeness (Danescu-Niculescu-Mizil et al. 2013). Beyond ad-hoc solutions, few tools exist that allow general purpose, combined semantic and syntactic analysis. Databases like FrameNet (Ruppenhofer et al. 2016) and PropBank (Kingsbury and Palmer 2002) codify relationships between words as annotated by experts (Gildea and Jurafsky 2002; Palmer, Titov, and Wu 2013). These approaches share the problem of data sparsity: there are many ways to convey a single message, and annotating all of them is difficult (Matsubayashi, Okazaki, and Tsujii 2010). In short, most combined semantic and syntactic analysis tools are either ad-hoc solutions, or require large data and significant effort to adopt. SENPAI complements these tools by identifying semantic and syntactic patterns in a corpus automatically.

SENPAI adds to recent efforts aiming to overcome the limitations of rich yet impractical framing analysis tools, such as automatically extracting lexico-syntactic patterns corresponding to personas (Card et al. 2016), narratives (Samory and Mitra 2018), and rhetorical devices (Zhang, Spirling, and Danescu-Niculescu-Mizil 2017).

4 SENPAI

SENPAI discovers combined syntactic and semantic patterns in a corpus. First, it groups together words with similar meaning—for example, it groups "@Obama" and "President Kennedy" because they both refer to individuals. Then, SENPAI identifies syntactic links that recur between similar words.

4.1 Finding semantically-related words

Certain groups of words have instrumental functions in sociolinguistic applications. For example, politeness assesses how hedges provide the addressee of a request with a face-saving way to deny it (e.g. "I **suggest** we start with...") (Danescu-Niculescu-Mizil et al. 2013). It is therefore important to identify word groups that are most appropriate for the task at hand. SENPAI uses a two-step approach to find groups of words with similar meaning. First, we adopt continuous word embeddings to measure semantic similarity between words (Pennington, Socher, and Manning 2014). Then, we use unsupervised artificial neural networks to group similar words together (Kohonen 1998).

Computing word similarity using word embeddings Specifically, we use spacy[3] to normalize words using lemmatization, so as to remove surface form variations which do not alter the meaning of a word, e.g. the lemma for both "moved" and "moves" is "move." Then, we encode lemmas with the corresponding 300-dimensional word vectors trained on the Common Crawl with GloVe (Pennington, Socher, and Manning 2014) (see figure 2, step 1).

Grouping similar words using self-organizing maps Next, we group together word embeddings with similar meaning (figure 2, step 2). To this end we use a self-organizing map (SOM), a neural network for dimensionality reduction (Kohonen 1998). SOMs are trained using competitive learning: several neurons compete for a word vector, and only the best matching neuron wins. We group together word vectors that activate the same neuron after training. Moreover, the neighborhood function in SOMs preserves the topological properties of the input. In other words, word vectors that are close in the original 300-dimensional space activate the same neuron. This way SOMs find clusters of words with a common abstraction (Kohonen 1998).

4.2 Discovering semantic and syntactic patterns

Syntax may drastically alter what semantics tell us. For example "he hit the club with me" and "he hit me with the club" have starkly different meaning despite being permutations of the same words—syntax marks the difference. Machines as well as humans need syntax to disambiguate sentences (Gildea and Palmer 2002; Levy and Goldberg 2014). How do we find consistent syntactic links between words of the same kind? SENPAI combines dependency parsing and graph mining to discover frequent syntactic relations.

Enhancing semantic clusters by analyzing syntax We use spacy to extract the syntactic dependency tree for words in the documents (figure 2, step 3). Then, we replace words with the SOM cluster they belong to. Thus, we represent documents as collections of rooted trees with syntactic dependencies as edge labels and with semantic clusters as node labels.

Finding common semantic and syntactic patterns Next, we identify patterns as parts of the trees that repeat in the corpus. To this end, we employ Gspan, a data mining algorithm that discovers frequent subgraphs efficiently (Xifeng Yan and Jiawei Han 2002). Gspan allows defining a minimum support $minsup$, i.e. a minimum number of times a pattern needs to appear in the forest

[3]https://spacy.io

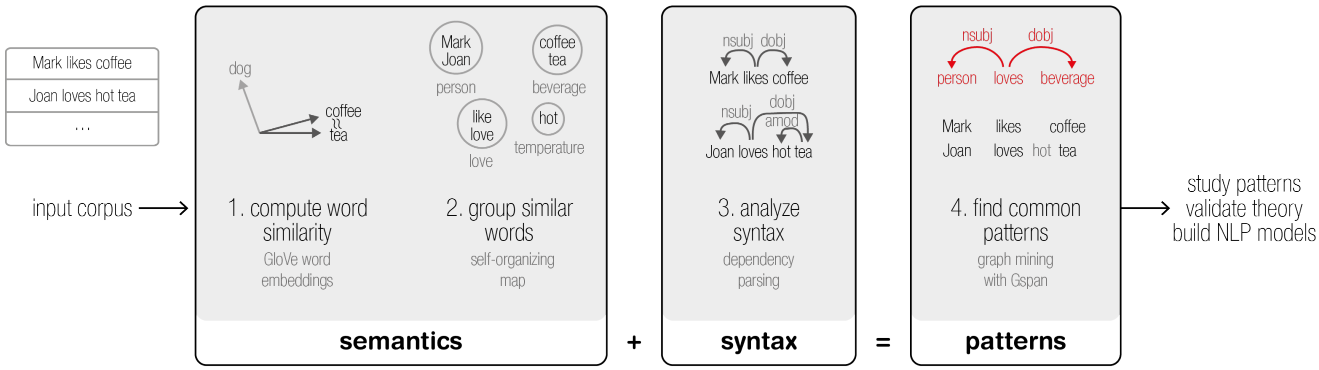

Figure 2: SENPAI discovers combined semantic and syntactic patterns from few example documents. It first encodes words using word embeddings, so that similar words are close in the semantic embedding space (step 1 in the figure). Then, it groups words into semantic clusters using a self-organizing map—a neural network that preserves semantic distance (step 2). In the example, "coffee" and "tea" are close in semantic embedding space, thus become grouped in the same semantic cluster. Further, SENPAI combines semantic cluster and syntactic link information (step 3). Finally, using approaches from data mining, SENPAI finds which of these patterns appear frequently in the example documents (step 4).

to be considered frequent, and a minimum number of nodes $minn$, i.e. a minimum number of words in the pattern for it to be considered interesting. In step 4 of figure 2, Gspan finds a pattern with nodes (semantic clusters) *person*, *loves*, *beverage*, where *person* and *beverage* are respectively the subject (*nsubj*) and object (*dobj*) of *loves* (syntactic links). A $minn$ of 3 allows discarding patterns that are too generic, e.g. "*person loves*," whereas a $minsup$ of 2 allows discarding patterns that are too rare, e.g. "*person loves temperature beverage*" in "Joan loves hot tea."

4.3 Selecting semantic and syntactic patterns

Depending on the application scenario, SENPAI also allows selecting patterns using alternative strategies.

Scenario 1: selecting patterns that best measure a linguistic dimension Ultimately, the patterns that SENPAI discovers may be useful for measuring a linguistic dimension, e.g. credibility, sentiment, or politeness. One way to select patterns that are most relevant for a particular dimension—the approach we use in this paper—is to use feature selection. We annotate the corpus to gather ground truth on the linguistic dimensions, e.g., by assessing the credibility, politeness, or sentiment of each document on a likert scale. In this paper, we use the crowdsourced annotations provided by the reference studies to demonstrate scenario 1. Similar to the bag-of-words approach in natural language processing, we then count how many times each pattern appears in a document. Then, we select patterns that are most significantly related to the ground truth annotations. To this end, we use a Lasso model with parameters determined in cross-validation[4]. A penalized model like Lasso allows discarding redundant and non-informative patterns.

Scenario 2: selecting patterns that characterize a corpus Whereas the previous approach concerns the variation of a linguistic dimension within a corpus, the current approach is useful when one is interested in which patterns distinguish the corpus itself. For example, one may want to explore how tweets by news media outlets differ from the typical tweet. In this case, a background corpus may serve as a contrast. We may gather a large yet unannotated corpus of tweets to represent the language on the platform. Then, we may select patterns that are over-represented in the news media tweets for example by measuring their log-odds ratios (Jurafsky et al. 2014).

Scenario 3: selecting patterns parsimoniously A third scenario presents itself when neither ground truth on the documents nor a background corpus is available. This scenario is typical when the goal is a purely exploratory analysis of the corpus. We may focus on a number of patterns that is small enough for human validation through the parameters of the graph mining algorithm. For example, we can decide the best trade-off between frequency ($minsup$) and complexity ($minn$) that results in no more than 100 patterns. A nonparametric alternative is to select the smallest set of patterns so that all documents show at least one pattern in the set. Although this is an instance of set cover, a known NP-complete problem, heuristic solutions exist and work well in practice (Ceria, Nobili, and Sassano 1998).

5 Qualitative evaluation

SENPAI can work as a black-box tool for quantifying linguistic dimensions by using its patterns as features in computational models. However, before evaluating how informative SENPAI's patterns are in computational models in section 6, we now assess if SENPAI can help humans gain insight on data. We test SENPAI on three social computing tasks: extracting patterns related to credibility, sentiment, and politeness in online text. To this end, we compare SENPAI to three corresponsing case studies: respectively (Soni et al. 2014), (Danescu-Niculescu-Mizil et al. 2013), and (Hutto and Gilbert 2014). The studies identified theoretically-motivated semantic and syntactic patterns, to measure the three linguistic dimensions. We use the case studies to qualify the interpretability, completeness, and meaningfulness

[4] Alternative approaches include ANOVA, SAGE, and CLES, among others

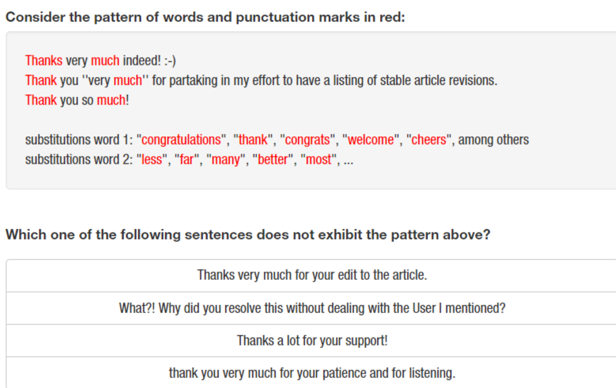

Figure 3: Depiction of the crowdsourcing task. We showed human annotators a pattern by highlighting it in red in three example sentences, and by providing alternative words that could substitute each one in the pattern. We then asked human annotators to recognize which out of four unseen sentences did not contain the pattern. In the figure, the pattern "thanks... much," in red in the upper, shaded section, is missing from the second sentence in the lower section.

of the patterns discovered by SENPAI. We focus on one dataset per linguistic dimension—news tweets for credibility, Wikipedia talk-pages for politeness, and Amazon reviews for sentiment, provided by the original papers. First, we show that human annotators without expert knowledge can correctly recognize the patterns discovered by SENPAI in unseen texts. Then, we show that SENPAI recovers patterns that theory considers meaningful for expressing the linguistic dimensions. Finally, we show how social computing researchers can use SENPAI to discover new patterns that theory did not account for.

5.1 Interpretability of SENPAI

Can users with no training in computational linguistics and no familiarity with the data, interpret SENPAI's patterns?

We asked human annotators on Amazon Mechanical Turk to infer the patterns from few examples, and to recognize them in unseen examples. In particular, we showed the human annotators three example documents with the pattern highlighted in red (see figure 3). To better convey the meaning of the words in the pattern, we also showed up to 5 substitute words for each word in the pattern. Previous research observed that offering example words and sentences helps disambiguate syntactic search queries (Muralidharan and Hearst 2014). Hence, we gave annotators four unannotated documents, of which exactly one did not contain the pattern. We asked the human annotators to identify the extraneous document, similarly to the "word intrusion" task introduced in (Chang et al. 2009). Figure 3 shows an example task: the upper shaded section, shows the common pattern present in three sentences (the pattern, in red, is "thank... much"); also in the shaded section are substitutions for the two words in the pattern, e.g. "congratulations" instead of "thanks." Finally, the lower section presents four sentences, of which exactly one (in this case the 2nd sentence) does not match the pattern and should be identified as the extraneous document.

task	dataset	# patterns	# correct	agreement
sentiment	Amazon	403	363	0.85
politeness	Wikipedia	228	184	0.75
credibility	news tweets	42	37	0.72
overall		673	584	0.81

Table 1: Human annotator correctly recognized 584/673 of patterns discovered by SENPAI—or 87%—across the three linguistic dimensions. We consider a pattern correctly recognized if at least 2 out of 3 human annotators gave the correct answer. Human annotators showed high inter-rater agreement, measured via Krippendorff's α.

In all, we asked human annotators to evaluate the 673 patterns, one pattern per Human Intelligence Tasks (HIT)—the unit of work in Amazon Mechanical Turk. We asked three human annotators to evaluate each pattern. To guarantee good English proficiency and high-quality responses, we limited to workers in the US, UK, and Canada, who had over 99% approval rate and had completed over 10,000 HITs. We further verified the fitness of the workers for the task through a qualification test. We asked workers to read the instructions and to correctly identify, in three separate instances, the extraneous sentence not matching a hand-curated pattern. We estimated through a trial run the average completion time of a HIT. According to this estimate we compensated human annotators \$0.12 per HIT, plus \$0.24–0.36 to account for the overhead for the mandatory qualification test, resulting in a fair hourly wage (Salehi et al. 2015).

Results: Table 1 summarizes the results. 87% of the patterns showed majority correct answers, Krippendorff's $\alpha = 0.81$. In particular, the percent of correctly identified patterns in each dataset was 88 for credibility ($\alpha = 0.72$), 81 for politeness ($\alpha = 0.85$), and 90 for sentiment analysis ($\alpha = 0.75$).

In a nutshell, human annotators were able to correctly identify the patterns suggested by SENPAI. This confirms that SENPAI extracts patterns that humans can recognize in the great majority of cases. Therefore, SENPAI could guide non-experts in annotating and validating semantic and syntactic patterns.

5.2 Comparing SENPAI-discovered patterns to theoretical patterns

How do patterns discovered by SENPAI compare to theoretically-motivated patterns? We reviewed the patterns that the case studies used to measure credibility, politeness, and sentiment, and compared them with those discovered by SENPAI. In particular, for each theoretically-motivated pattern, we looked for a matching pattern among the ones discovered by SENPAI. All of the theoretically-motivated patterns matched a pattern automatically discovered by SENPAI (see table 2), with a single exception.

Case study 1: Credibility. Soni et al. measured the credibility of journalistic reports on Twitter using two strategies (Soni et al. 2014). First, they used validated dictionaries to find *cue* verbs, like "say", "report", and "tell", that signal the presence of a reported claim (Sauri 2008). The authors used dependency relations to identify the source of the claim and the claim itself. In a second strategy, they used regular expressions to find sentences using the pattern "claim*, according to* source". SENPAI recovered patterns induced by both cue verbs and the formulaic "according to" sentence structure.

Case study 2: Politeness. Danescu-Niculescu-Mizil et al. studied semantic and syntactic patterns in polite requests. The authors relied on a theoretical understanding of language to identify expressions related to language, like deference and gratitude. As for the credibility case study, all theoretically-inspired politeness patterns matched at least one pattern discovered by SENPAI.

Case study 3: Sentiment. Hutto and Gilbert developed VADER, a sentiment analysis tool combining a gold-standard lexicon with grammatical and syntactical patterns, including for example the use of exclamation points or degree modifiers (Hutto and Gilbert 2014). The patterns discovered by SENPAI contained all but one of the theoretically-motivated lexicons and rules. The only rule in VADER that SENPAI was not able to capture was the use of all-caps to emphasize words, because word embeddings ignore formatting.

In sum, SENPAI automatically discovered most of the linguistic patterns that theory indicated as meaningful. Researchers could therefore use SENPAI to validate the presence of theoretically-motivated patterns.

5.3 Using SENPAI to discover new patterns

We showed that *some* of the patterns discovered by SENPAI are known indicators of the linguistic dimensions under study. One might ask if the remaining patterns are just technically informative—after all they have nonzero Lasso coefficients, see section 4.3—or if they actually provide new, meaningful measures of the linguistic dimensions under study. To assess the meaningfulness of those patterns, we selected the 20 most representative patterns from each linguistic dimension according to their Lasso coefficients. Then, we showed the selected patterns to two graduate researchers who had domain knowledge about the three case studies but were not NLP experts. We provided them with three documents containing each pattern highlighted in bold. Moreover, for each word in the pattern, we showed the researchers up to five words in the same semantic cluster (similar to the task shown in figure 3). We also told them whether the patterns had positive or negative relations with the linguistic dimensions—e.g. appearing in polite or impolite requests in the Wikipedia dataset. We asked each researcher independently: 1) if the patterns exhibited recognizable semantic and syntactic properties; 2) if those properties could explain the relation between the patterns and the linguistic dimensions. We then met in person to discuss their responses. We focused on new patterns that were not explainable by the theoretical assumptions in the original papers, and we reached consensus on their properties. We report the new patterns that SENPAI discovered in table 3.

Case study 1: Credibility. SENPAI expanded on the patterns suggested by theory. For example, not only SENPAI's patterns contained the cue verbs in the validated dictionaries like *confirm, observe, predict, suggest*, SENPAI also suggested other verbs that journalists used in the same fashion, e.g. *assure*, *mention*, and *clarify*. Moreover, SENPAI discovered patterns that are new altogether. For example, it highlighted a specific use of testimonials: citing people especially clarifying their official

recovering theoretically-motivated patterns

theory	***matching pattern from SENPAI***
credibility of journalistic tweets (Soni et al. 2014)	
Cue-verb-induction	RT @Edgecliffe: **Graham Holdings Company says** it's selling Washington Post HQ building to Carr Properties
According-to-induction	Howard Zinn had a disappointingly simplistic and ideological view of history **according to** David Greenberg @tnr http://t.co/jJd19rAsjB
politeness of Wiki talk (Danescu-Niculescu-Mizil et al. 2013)	
Gratitude	**Thanks** very much **for** your edit to the article.
Deference	**Great work** you've been doing on this.
Greeting	**Hi.** How do you think, is of FL quality?
Positive lexicon	Otherwise, **very nice** article.
Negative lexicon	What **the hell**?
Apologizing	**Sorry for** not replying sooner.
Please	Hey, can **you please tell** me if has been fixed enough by being stub-ified?
Please start	**Please** could **you tell** me where I can find page-views for wiki articles?
Indirect (btw)	**By** the **way**, are you honestly ok with me bothering you like this?
Direct question	**What is** the source of this image?
Direct start	**Or are** you an incurable masochist**?**
Counterfactual modal	On a tangential note, **would** you **happen** to be?
Indicative modal	**Can** you **fix** that, please?
1st person start	**I believe** you wanted to find out which party the PP was?
1st person pl.	If **we**'re **going** to delete every fair use image ...
1st person	Did **I mention** that she was in?
2nd person	How come **you**'re **taking** on a third language**?**
2nd person start	**You owe** me one back?
Hedges	**Not** too **sure** now you come to mention it.
Factuality	Can I ask: **was** this **really** the implication that you intended in making this change?
sentiment of Amazon reviews (Hutto and Gilbert 2014)	
Positive lexicon	sound quality is **amazing.**
Negative lexicon	the scroll wheel, to be blunt, **sucks.**
Exclamation point	i **love** this phone**!!!**
Degree modifiers	this camera is **insanely great**!
"But" shifts sentiment	[...] **has** good sound volume **but** it hurts the ears
Negation inverts sentiment	the other thing **i don't like** [...] / i have **no problems** with it.

Table 2: SENPAI discovers patterns that theory suggests are important for measuring the linguistic dimensions under study—credibility, politeness, and sentiment. For each theoretically-motivated rule in the original studies (Soni et al. 2014; Danescu-Niculescu-Mizil et al. 2013; Hutto and Gilbert 2014), we report in bold an example pattern discovered automatically by SENPAI that matches the rule.

discovering new patterns	
new theory	***matching pattern from SENPAI***
credibility of journalistic tweets (Soni et al. 2014)	
Testimonials +	In response to French reporter **, Pres Obama says** he would not choose between his daughters & he won't choose between US alllies in Europe.
Negations -	RT @joshrogin: Obama **didn't think** Congress had to weigh in before he decided to intervene in Libya #justsayin
politeness of Wiki talk (Danescu-Niculescu-Mizil et al. 2013)	
Asking for feedback +	**What do** you **think**?
Asking permission +	**Is** that **OK** with you?
Asking permission +	Would **you mind** if I copied and pasted your version over to mine to work on?
Rhetorical questions -	Correct me if I am wrong, but **you don't seem** to have asked if anyone had a problem with the merge, did you?
Rhetorical questions -	Anyway **why put** up with all this carelessness?
sentiment of Amazon reviews (Hutto and Gilbert 2014)	
Endorsement +	overall, i **highly recommend** this phone.
Value for money +	terrific mp3 player, **great price**.
Cost -	needless to say, i wouldn't recommend anyone **purchasing** this **product**.
Cost -	why **does** a 256 mb player **cost** $$$
Returns -	i had hoped this was bought from walmart so that **i** could **return** it!

Table 3: SENPAI discovers new patterns from data that intuitively and quantitatively measure the three linguistic dimensions credibility, politeness, and sentiment. We mark new patterns positively and negatively related to credibility with the symbols + and -. This table reports only few examples that go beyond the theories in the case studies: these patterns would therefore be missed adopting a theoretical approach.

title, e.g. by denoting the presidential title in "**Pres Obama**". Precise contextual details like official titles boosts the perceived factuality (Liao and Shi 2013). Whereas testimonials increase credibility, phrasing statements as negations decreases it, e.g. "Obama **didn't think**..."

Case study 2: Politeness. SENPAI discovered several new patterns that relate to politeness. Asking for feedback ("**what do** you **think**...") and for permission ("**Is** that **OK**...") correspond to polite requests. These strategies try to build a positive relationship between the requester and the addressee—similarly to the greetings and deference patterns in the original paper (Danescu-Niculescu-Mizil et al. 2013), but missed in the original paper which relied only on theory and not on data. Furthermore, the annotators interpreted rhetorical questions (e.g. "**why put** up with..." and "**you** do**n't seem** to have asked...') as a passive-aggressive way of phrasing requests—indeed, those patterns correspond to impolite requests, which may therefore inform new theory.

Case study 3: Sentiment. SENPAI surfaced new patterns that strongly relate to human assessments of sentiment. Reviews were mostly negative when they mentioned dealing with the sellers' customer service and returning items. Reviews were also negative when they mentioned the cost of the items ("why **does** it **cost** ..."). Yet, products received better reviews when they proved good value for the money ("**great price**"), and when users showed their endorsement ("**highly recommended**"). These different factors show how sentiment is not unidimensional—i.e., buyers do not just consider the product's price point but also its return on investment. In fact, research directions in sentiment analysis include capturing sentiment towards specific (Lakkaraju, Socher, and Manning 2014) or contrastive (Fang et al. 2012) aspects of individuals' opinions. These new, data-driven patterns would be missed by a purely theoretical approach.

In sum, researchers could use SENPAI to generate candidate linguistic patterns directly from data, without the selection bias that comes with assuming which patterns should be relevant, and without necessarily being NLP experts.

6 Quantitative evaluation

The three case studies used theory-driven semantic and syntactic patterns to build computational models, and evaluated the performance of the models using crowdsourcing. Next, we show that substituting theoretically-motivated patterns with patterns automatically discovered by SENPAI results in better computational models. We tested model performance on seven different datasets across the three case studies. Table 4 summarizes the performance of models using SENPAI and compares it to the corresponding models from the original papers. Lastly, we show that refining SENPAI through the feedback by crowdworkes yields interpretable features without loss in predictive power.

6.1 Case study 1: credibility of online news

Soni et al. measured the credibility of news tweets (Soni et al. 2014)[5]. They focused on specific patterns in the tweets, such as the presence of verbs like "confirms" and "predicts," and whether the subjects of the verbs are named entities like "Obama" or "@abc." Then, they used these theoretically-motivated patterns as features in a ridge regression model, and trained the model on the credibility annotations of the tweets. The authors evaluated the performance of the model using its mean average error (MAE, where lower MAE means better performance). They also used a baseline model which always predicted the mean credibility in the training set. They compared the model against the baseline using z-test. We used SENPAI to automatically discover patterns. Then, we replicated the authors' model design with features replaced with the ones discovered by SENPAI.

Results: Models using SENPAI not only outperformed the baseline ($MAE = 0.366$ vs. $MAE = 0.425$, z-test $p < 0.001$) but also the model using cue words ($MAE = 0.376$, $p < 0.001$)—the best performing theoretically-motivated model. Moreover, the model using SENPAI outperformed a model using all other features combined ($MAE = 0.391$, $p < 0.001$). Adding patterns discovered by SENPAI to the latter model resulted in increased performance ($MAE = 0.381$, $p < 0.001$).

[5]https://github.com/jacobeisenstein/twitter-certainty

task	metric	SENPAI	theory-inspired model	base-line	dataset
credibility	MAE	**0.36**	0.37	0.42	Twitter†
politeness	ACC	**0.83**	0.80	0.79	Wikipedia†
		0.65	0.63	0.63	SE
sentiment	F1	**0.67**	0.60	0.62	Amazon†
		0.79	**0.91**	0.77	Twitter
		0.75	0.58	0.59	IMDB
		0.54	0.54	0.52	NYT

Table 4: Computational models using SENPAI perform better than theoretically-inspired models in measuring different linguistic dimensions and in diverse datasets. We highlight in bold the best performing models. We complement quantitative findings with qualitative assessments on datasets marked with †.

This suggests that SENPAI captures as much information as the models in the original paper, if not more.

6.2 Case study 2: politeness of online requests

Danescu-Niculescu-Mizil et al. studied the politeness of requests in two online domains: Wikipedia talk-pages and StackExchange (Danescu-Niculescu-Mizil et al. 2013)[6]. They trained a linear SVM model to predict the politeness score using 20 theoretically-motivated semantic and syntactic patterns, together with all case-insensitive unigrams appearing at least 10 times in the data. The model was evaluated on its accuracy in leave-one-out cross-validation. The authors also trained a baseline model using only unigrams, i.e. excluding the 20 theoretically-motivated patterns. We used SENPAI to discover patterns, and substituted them for the theoretically-motivated patterns in the politeness model.

Results: The politeness classifier scored 80.1% accuracy on the Wikipedia dataset. The classifier powered by SENPAI scored 83.1% accuracy, better than both the baseline and the original model in a statistically significant way ($p < 0.05$) and close to the human performance of 86% (as reported in the original paper). The StackExchange dataset yielded similar results. The original politeness classifier scored 63% accuracy, whereas the SENPAI-powered classifier 65% ($p<0.05$). The reference study considered Wikipedia its development domain because it was used for identifying and defining patterns. Whereas in the previous case study Soni et al. mostly relied on theoretical expertise for defining the patterns for measuring credibility, Danescu-Niculescu-Mizil et al. gained domain insight from data. This corroborates that SENPAI performs favorably to expertise in computational linguistics as well as domain knowledge.

6.3 Case study 3: sentiment of online text

VADER uses a combination of validated lexicons and syntax-aware rules to measure sentiment (Hutto and Gilbert 2014). They validated VADER on four datasets: social media texts from Twitter, movie reviews from IMDB, reviews of technological products from Amazon, and opinion editorials from the New York Times[7]. The authors measured VADER's performance by using it to categorize texts into three classes—positive, neutral, and negative—and then computing the F1 score with respect to the ground truth annotations. We could not integrate SENPAI directly into VADER because the latter uses a custom algorithm to attribute weights to words in a sentence and to combine them into a single sentiment score. Instead, we used a Naive Bayes classifier to learn the weights of the patterns and to combine the weights into a three-class prediction in 100-fold cross-validation. As a baseline, we also compare SENPAI to the second-best results reported in (Hutto and Gilbert 2014), namely the ones using the approach in (Hu and Liu 2004).

Results: The classifier using SENPAI scored better than VADER on the Amazon reviews ($F1=0.67$ vs. 0.60) and movie reviews (0.75 vs. 0.58) dataset. VADER, conversely, performed better on tweets (0.91 vs. 0.79). VADER was specifically tuned to measure sentiment in social media using this dataset, and it performed better than the human raters who annotated it ($F1 = 0.84$). Despite VADER's outstanding performance, SENPAI still scored close to human raters, and better than all other lexicons VADER compared itself against (second-best $F1=0.77$ (Hutto and Gilbert 2014)). Finally, the SENPAI-based classifier and VADER performed similarly on New York Times editorials ($F1 = 0.54$ for both). Measuring sentiment in news is admittedly difficult, because news aim to be objective and scarcely use emotional language. The fact that even human raters performed modestly on this dataset ($F1 = 0.65$) may explain the small margin between SENPAI and VADER.

6.4 Limiting features to human-validated patterns

Finally, we tested the performance of the models using only the patterns that the human annotators were able to interpret as features (see section 5.1 for a summary of the crowdwork task). The accuracy of the models remained high. The credibility model showed $MAE = 0.380$, better than the baseline ($p < 0.001$). The politeness model scored 83.1% accuracy, like the model including non-interpretable patterns. The sentiment models scored 67% F1, similar to the model including all patterns. In other words, we can use SENPAI in combination with crowdsourcing to obtain interpretable models without loss in predictive power.

7 Discussion

Computational models powered by SENPAI match the performance of theoretically-motivated ones. SENPAI does so without any prior knowledge of linguistics nor the application domain. Therefore, SENPAI could enable users with no expertise in natural language processing to gain insight on high-level linguistic properties of corpora, such as sentiment, credibility, and politeness. In fact, we found that users without expertise in computational linguistics correctly interpret the patterns suggested by SENPAI and recognize them in new documents—social computing researchers who had domain knowledge but lacked NLP expertise. Furthermore, patterns surfaced by SENPAI appear relevant: in all three case studies, SENPAI discovered patterns that included the theoretically-motivated ones proposed by experts. Thus, computational linguists could use SENPAI to validate theoretical assumptions, such as which patterns and conventions suggested by theory are present in the data. Moreover, since SENPAI is data-driven, it allows

[6] https://www.cs.cornell.edu/~cristian/Politeness.html
[7] https://github.com/cjhutto/vaderSentiment

researchers to bypass such assumptions altogether and to identify representative patterns and to derive theories directly from data.

7.1 Practical implications

We next discuss how researchers may use SENPAI.

Discover patterns that work without expertise SENPAI automates finding which textual patterns occur in the data, and which of those patterns are best suited to measure a linguistic dimension, like sentiment or credibility. Researchers interested in measuring linguistic dimensions can use SENPAI to build computational models. By leveraging approaches from NLP and data mining, SENPAI automatically discovers relevant patterns in a corpus, and encodes the corpus as a bag-of-patterns by identifying which patterns are present in each document. Researchers can therefore use SENPAI as a drop-in replacement for lexicon and bag-of-words encoders. SENPAI offers the advantage over the latter because it encodes richer information by combining semantic and syntactic patterns instead of just counting words. Furthermore, we showed how researchers can also identify patterns that best predict the linguistic dimensions of interest through crowdsourcing. In other words, SENPAI allows studying socio- and psycho-linguistic dimensions by removing the need for expert theoretical, domain-specific, and NLP knowledge.

Gain insight from data to validate theory We showed how experts can use SENPAI to validate the presence of theoretically-motivated patterns as well as to surface application domain-specific patterns. This process can confirm theoretical assumptions about the use of language, or conversely can identify theoretical blind spots. Linguistics theories developed on scholarly English may not apply to social media texts—e.g. because of the use of hashtags, mentions, and emojis (Barbieri et al. 2018; Pavalanathan and Eisenstein 2015; Stewart et al. 2017). Since SENPAI identifies relevant patterns from scratch, it only depends on the choice of the example documents to analyze. Thus, computational linguists can extract data-driven patterns using SENPAI and code them so as to reconstruct the theoretically-motivated patterns. Therefore, computational linguists can use SENPAI to validate previous theories on new application domains before expanding on the theories themselves. This arguably accelerates the research process.

8 Limitations and future work

SENPAI is not without limitations, and researchers should take them into consideration when using SENPAI. One of the strengths of SENPAI is that it depends solely on selected examples. This helps mitigate selection bias for some patterns over others. The reverse of the medal is that SENPAI is not immune from sampling bias. SENPAI does not model general properties of language, but rather identifies the qualities of an application context. For example, the study by Soni et al. focuses on news reports on social media: SENPAI identifies patterns that characterize journalistic style in twitter conversations, but that would not be applicable to newspaper articles. Another limitation is that SENPAI uses under the hood general purpose word embeddings and dependency parsing tools. Arguably, training the underlying models on data from the same domain as the input examples would yield better overall accuracy—and indeed it is possible whenever in-domain data is available.

We believe that SENPAI would be the most useful when combined with expert human judgment. SENPAI-powered interactive tools would enable at one time to explore data effectively and to quickly iterate on refining language models. Quickly adapting linguistic models to new patterns is a much needed capability, e.g. in the cat-and-mouse chase between spammers and spam filters or abusers and automated moderators. Doting SENPAI with a user interface would also help curating and editing the patterns. For example, it would make it easy to annotate, merge, split, delete, edit, and add patterns.

9 Conclusions

This paper introduced SENPAI, an interpretable tool for exploratory language analysis. It discovers combined semantic and syntactic patterns from a few examples. Researchers can use SENPAI as a drop-in replacement for lexicons or bag-of-words encoders, with the additional benefit of richer feature representation. SENPAI is, first, as useful as theoretically-motivated patterns in measuring high-level linguistic dimensions. Models using SENPAI consistently outperformed models using expert-curated patterns on three widespread computational linguistic tasks in three different application domains. Second, SENPAI is interpretable: users without particular expertise can inspect patterns directly from examples. Human annotators were able to identify 87% of the patterns correctly in new documents. Third, SENPAI is meaningful, in that experts can use it to validate theoretically-motivated assumptions and to better frame them into specific application domains.

References

Barbieri, F.; Marujo, L.; Karuturi, P.; and Brendel, W. 2018. Multi-task Emoji Learning. In *1st International Workshop on Emoji Understanding and Applications in Social Media*.

Baronchelli, A. 2018. The emergence of consensus: a primer. *Royal Society Open Science* 5(2):172189.

Cambria, E., and White, B. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9(2):48–57.

Card, D.; Gross, J. H.; Boydstun, A. E.; and Smith, N. A. 2016. Analyzing Framing through the Casts of Characters in the News. *EMNLP* 1410–1420.

Ceria, S.; Nobili, P.; and Sassano, A. 1998. A Lagrangian-based heuristic for large-scale set covering problems. *Mathematical Programming* 81(2):215–228.

Chancellor, S.; Hu, A.; and De Choudhury, M. 2018. Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities. In *CHI*, 1–14.

Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS*, 288–296.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. *Proceedings of the ACL* (July):892–901.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. *Proceedings of the ACL*.

Donath, J. 1999. Idenity and Deception in the Virtual Community. *Communities in Cyberspace* 27–58.

Eisenstein, J. 2018. Identifying Regional Dialects in On-Line Social Media. *The Handbook of Dialectology* 2013(December 2013):368–383.

Ellis, N. C. 2002. Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition* 24(02):143–188.

Fang, Y.; Si, L.; Somasundaram, N.; and Yu, Z. 2012. Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model Categories and Subject Descriptors. *WSDM* 63–72.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding Topic Signals in Large-Scale Text. *CHI* 4647–4657.

Garimella, K.; Weber, I.; and De Choudhury, M. 2016. Quote RTs on Twitter. In *WebSci*, 200–204.

Gilbert, E. 2012. Phrases that signal workplace hierarchy. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12* 1037.

Gildea, D., and Jurafsky, D. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3).

Gildea, D., and Palmer, M. 2002. The necessity of parsing for predicate argument recognition. In *Proceedings of the ACL*, 239–246.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. *KDD* 168.

Hutto, C. J., and Gilbert, E. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *ICWSM*, 216–225.

Jurafsky, D.; Chahuneau, V.; Routledge, B. R.; and Smith, N. A. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19(4).

Kasunic, A., and Kaufman, G. 2018. "At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. *ICWSM*.

Kingsbury, P., and Palmer, M. 2002. From TreeBank to PropBank. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* 1989–1993.

Kohonen, T. 1998. The self-organising map. *Neurocomputing* 21(1-3):1–6.

Lakkaraju, H.; Socher, R.; and Manning, C. D. 2014. Aspect Specific Sentiment Analysis using Hierarchical Deep Learning. *NIPS WS on deep neural networks and representation learning*.

Levy, O., and Goldberg, Y. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *CoNLL*, 171–180.

Liao, Q., and Shi, L. 2013. She gets a sports car from our donation: Rumor Transmission in a Chinese Microblogging Community. In *CSCW*.

Matsubayashi, Y.; Okazaki, N.; and Tsujii, J. 2010. Generalization of Semantic Roles in Automatic Semantic Role Labeling. *Journal of Natural Language Processing* 17(4):59–89.

Muralidharan, A., and Hearst, M. A. 2014. Improving the Recognizability of Syntactic Relations Using Contextualized Examples. In *Proceedings of the ACL (Short Papers)*, 272–277.

Nguyen, D.; Doğruöz, A. S.; Rosé, C. P.; and de Jong, F. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics* 42(3):537–593.

Nguyen, D.; McGillivray, B.; and Yasseri, T. 2017. Emo, Love, and God: Making Sense of Urban Dictionary, a Crowd-Sourced Online Dictionary.

Palmer, M.; Titov, I.; and Wu, S. 2013. Semantic Role Labeling. *NAACL HLT 2013 Tutorial Abstracts* 10–12.

Pavalanathan, U., and Eisenstein, J. 2015. Audience-Modulated Variation in Online Social Media. *American Speech* 90(2):187–213.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M. R. L.; Johnson, C. R.; Baker, C. F.; and Scheffczyk, J. 2016. *FrameNet II: Extended Theory and Practice*.

Salehi, N.; Irani, L. C.; Bernstein, M. S.; Alkhatib, A.; Ogbe, E.; Milland, K.; and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *CHI*, 1621–1630.

Samory, M., and Mitra, T. 2018. "The Government Spies Using Our Webcams": The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM* CSCW.

Sauri, R. 2008. *A factuality profiler for eventualities in text*. Ph.D. Dissertation.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8(9).

Soni, S.; Mitra, T.; Gilbert, E.; and Eisenstein, J. 2014. Modeling Factuality Judgments in Social Media Text. In *Proceedings of ACL*, 415–420.

Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3):538–556.

Stewart, I.; Chancellor, S.; De Choudhury, M.; and Eisenstein, J. 2017. #anorexia, #anarexia, #anarexyia: Characterizing Online Community Practices with Orthographic Variation. In *SocialNLP*, 1–9.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *WWW*.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.

Xifeng Yan, and Jiawei Han. 2002. gSpan: graph-based substructure pattern mining. In *ICDM*, 721–724.

Zhang, J.; Spirling, A.; and Danescu-Niculescu-Mizil, C. 2017. Asking too much? The rhetorical role of questions in political discourse. In *EMNLP*.