Can you read my mindprint? Automatically identifying mental states from language text using deeper linguistic features

Lisa S. Pearl
Department of Cognitive Sciences
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
lpearl@uci.edu

Igii Enverga
Department of Computer Science
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697
envergaj@uci.edu

Abstract

Humans routinely transmit and interpret subtle information about their mental states through the language they use, even when only the language text is available. This suggests humans can utilize the linguistic signature of a mental state (its *mindprint*), comprised of features in the text. Once the relevant features are identified, mindprints can be used to automatically identify mental states communicated via language. We focus on the mindprints of eight mental states resulting from intentions, attitudes, and emotions, and present a mindprint-based machine learning technique to automatically identify these mental states in realistic language data. By using linguistic features that leverage available semantic, syntactic, and valence information, our approach achieves near-human performance on average and even exceeds human performance on occasion. Given this, we believe mindprints could be very valuable for intelligent systems interacting linguistically with humans.

Keywords: mental state, linguistic features, mindprint, natural language processing, information extraction

Biographical Note

Lisa Pearl is an Associate Professor of Cognitive Sciences, Linguistics, and Logic & Philosophy of Science at the University of California, Irvine, and directs the Computation of Language laboratory. Igii Enverga is a junior undergraduate student in Computer Science at the University of California, Irvine and a research assistant in the Computation of Language laboratory. His interests include natural language processing and artificial intelligence.

1 Introduction

Humans routinely communicate subtle information about their mental states through the language they use. Often, they are fairly good at transmitting and interpreting this information, which is usually perceived as a message's tone, even when only the language text is available (Kruger, Epley, Parker, & Ng, 2005; Pearl & Steyvers, 2010, 2013). While there is certainly individual variation in this ability (e.g., Pearl & Steyvers, 2010), humans typically are sufficiently competent at it that a diagnostic feature of disorders such as Asperger's Syndrome is the miscomprehension of message nuance (McPartland & Klin, 2006).

This indicates that, in the absence of auditory and visual cues, humans can express certain aspects of their mental states in addition to the basic semantic content of their messages. For example, in the message "Don't you just love this idea?", the basic semantic content is something like *love(you, this idea)*, while the tone is persuasive or possibly sarcastic, expressing the speaker's intention to persuade the listener of the content (persuasion) or perhaps communicate the opposite of the content in an amusing way (sarcasm).

This suggests that there are linguistic signatures for mental states – which we call *mindprints* – comprised of features in the language text, and humans are capable of utilizing these mindprints to express and perceive different mental states. Once the relevant linguistic features of mindprints are identified, mindprints can be used to automatically identify the mental states underlying messages more generally. Because of this potential, much recent research has focused on leveraging linguistic features to automatically identify mental states such as *intentions* (Mihalcea & Strapparava, 2009; Pearl & Steyvers, 2010; Anand et al., 2011; Rubin & Conroy, 2011; Pearl & Steyvers, 2013), *attitudes* (Mishne, 2005; Keshtkar & Inkpen, 2009; Neviarouskaya, Prendinger, & Ishizuka, 2010; Pearl & Steyvers, 2010; Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, & Potts, 2013; Pearl & Steyvers, 2013), *emotions* (Mishne, 2005; Strapparava & Mihalcea, 2008; Keshtkar & Inkpen, 2009; Neviarouskaya et al., 2010; Pearl & Steyvers, 2010; Chaffar & Inkpen, 2011; Mohammed, 2012; Pearl & Steyvers, 2013), and *perspectives* (Lin, Wilson, Wiebe, & Hauptmann, 2006; Hardisty, Boyd-Graber, & Resnik, 2010).

We first review previous approaches to mental state identification that are related to the approach we pursue here. Notably, approaches have often been developed that target a particular mental state (e.g., politeness: Danescu-Niculescu-Mizil et al., 2013) or mental state type (e.g., emotions: Chaffar & Inkpen, 2011; Mohammed, 2012), rather than mental state identification more generally. There are, of course, numerous possible mental states – for example, the blogging site Livejournal suggests over a hundred potential moods to its users (Mishne, 2005; Keshtkar & Inkpen, 2009). However, previous approaches have tended to target specific subsets of mental states because additional domain-specific knowledge has been leveraged to identify the mental state(s) of interest (e.g., hedges in politeness: Danescu-Niculescu-Mizil et al., 2013; affect lexicon features for emotions: Chaffar & Inkpen, 2011; Mohammed, 2012). A common first step in developing an automated approach is to identify which mental state is present from a circumscribed list of mental states (e.g., Chaffar and Inkpen (2011) identify which of six emotions is present) or to assign a score along a scale relating to one mental state (e.g., Danescu-Niculescu-Mizil et al. (2013) generate a score for

a message ranging from very polite to very impolite).

Still, recent approaches for identifying different types of mental states have been successful when using only shallow linguistic features (Keshtkar & Inkpen, 2009; Pearl & Steyvers, 2013). Here, we focus on mindprints that incorporate more sophisticated linguistic features involving semantic, syntactic, and valence information, but notably no explicit domain-specific knowledge about a mental state. That is, these linguistic features are not targeted for a specific mental state or mental state type the way that a hedge feature for politeness would be, or affect features for emotions would be. We apply these to the identification of eight mental states spanning intentions, attitudes, and emotions. Our study thus continues the investigation of this more difficult mental state identification task, where the possible mental states are not all of the same type, and we focus on the linguistic features that can be utilized to solve it.

After we describe the linguistic features allowed in our mindprints, we discuss the language dataset we use to specify the features for each mental state. This dataset is drawn from language data generated by the Word Sleuth game-with-a-purpose (*wordsleuth.ss.uci.edu*), which was specifically designed by Pearl and Steyvers (2013) to create a database that can be used for mental state identification research. In this game, human players are encouraged to both create messages expressing a specific mental state and interpret messages previously created by other players. Thus, these data have two useful properties. First, they indicate the mental state intended for a particular message, and so provide a known ground truth for every message. Second, they indicate which mental state was perceived in a particular message, and so provide a metric of human performance on this task. The performance of our automated mindprint-based techniques can thus be easily measured against human ability to detect the intended mental state in these messages.

We find that machine learning classifiers using more linguistically sophisticated mindprints perform very well: they achieve near-human level performance on average and even exceed human performance in one case. In addition, they significantly outperform classifiers using mindprints that are based only on shallow linguistic features, underscoring the utility of deeper linguistic features. Interestingly, we find that our best classifiers make similar errors to humans in some cases, suggesting that the mindprints used here are similar to the mindprints human use. Given the promising results we have found, we conclude that linguistically-sophisticated mindprints are likely to be a viable and important component of any intelligent system interacting linguistically with humans.

2 Related approaches

While much research has focused on leveraging linguistic features associated with a specific mental state (e.g., politeness) or mental state type (e.g., emotions), few studies have attempted to identify the linguistic features that could comprise mindprints more generally. This may be because researchers have assumed (not unreasonably) that much of how humans transmit information about a mental state involves knowledge that pertains only to that mental state. For example, to convey politeness, strategies such as asserting common ground and avoiding

disagreement can be used (Brown & Levinson, 1987), while to convey a particular emotion (e.g., *anger*), lexicon items that are associated with that emotion (e.g., *mad*) can be used.

Nonetheless, some research on the identification of mental states more generally has examined the efficacy of shallow linguistic features. Research in the classification of moods, (Mishne, 2005; Keshtkar & Inkpen, 2009), which span emotions such as *happy*, attitudes such as *contemplative*, and physically-based states such as *sleepy*, has found mixed results. Mishne (2005) combined support vector machines (SVMs) with shallow linguistic features (see Table 1) to identify 40 moods in Livejournal blog text, achieving performance barely above baseline (57% on average with a 50% baseline[1]). Notably, Mishne (2005) found that humans struggled to identify the moods in the linguistic data used for evaluation, achieving only 63% accuracy (again with a 50% baseline). Thus, these particular data may not have been very reliable to begin with for the classification task chosen.

Still, Keshtkar and Inkpen (2009) used this same dataset and attempted to identify 132 moods using a subset of the features Mishne (2005) used (see Table 1). Notably, they employed a more intricate machine learning technique involving a cascading set of SVMs, with each SVM classifying a mood into a finer-grained set of possible moods (e.g., a sequence of classifications for a mood might first be *sad*, then *uncomfortable*, given that it was *sad*, and then *cold*, given that it was *uncomfortable*). Their cascading approach achieved 55% accuracy (with a 7% baseline), which suggests that fairly shallow linguistic features can be more effective if more sophisticated machine learning techniques are utilized.

In this vein, Pearl and Steyvers (2013) investigated techniques to identify eight mental states spanning intentions, attitudes, and emotions that involved additional shallow linguistic features (see Table 1) and a different sophisticated machine learning algorithm. Their language data consisted of a set of brief messages (average length = 11.5 words) generated in the Word Sleuth game-with-a-purpose. Due to the design of the game, the messages used for their classification research tended to be more reliable with respect to mental state transmission, with humans successfully interpreting the intended mental state in a message 74% of the time on average (given a random guessing baseline of 12.5%). Pearl and Steyvers (2013) employed the Sparse Multinomial Logistic Regression (SMLR) classifier (Krishnapuram, Figueiredo, Carin, & Hartemink, 2005), which simultaneously learned (i) the linguistic features that were most useful for identifying each mental state and (ii) which mental state was being expressed in each message. Thus, the SMLR classifier identified the features comprising the mindprint for each mental state at the same time as it learned to identify the mental state with that mindprint. This technique achieved 70% accuracy on average (with the same baseline as humans had of 12.5%), and identified promising mindprint features for all the mental states. Given this success, it seems that machine learning approaches leveraging mindprints have the potential to identify mental states in language text as well as humans do by paying attention to the kind of linguistic features that humans may use. It is with this in mind that we pursue a

[1]The basline was 50% since the classification task was to indicate whether the text was an example of the mood or not. For example, a text would be rated as *happy* or *not happy*, and the perfect outcome would be for all *happy* texts to be labeled *happy* and all other texts to be labeled *not happy*. This classification was conducted for each mood, so the same texts would then be rated as *contemplative* or *not contemplative*, and then as *sleepy* or *not sleepy*, and so on.

richer set of linguistic features from which to build mindprints for mental states.

Table 1: Linguistic feature types used by related approaches: M2005 = Mishne (2005), K&I2009 = Keshtkar & Inkpen (2009), P&S2013 = Pearl & Steyvers (2013). Two classes of features are represented. The first class is drawn from standard features used in stylometric analysis and potentially reveals subconscious linguistic style changes caused by different mental states (parts of speech, length, punctuation, characters, lexical diversity, 1^{st} person pronouns, average word log frequency). The second class is more directly linked to particular mental states, as it often can capture specific items or expressions related to a mental state (n-grams, point-wise mutual information, valence, emphasized words, emoticons).

Feature type	**Examples**	**M2005**	**K&I2009**	**P&S2013**
parts of speech	singular noun determiner	✓		
length	5.2 words per sentence 30 word tokens total	✓	✓	✓
punctuation	... !	✓		✓
characters	all digits *p*			✓
lexical diversity	$\frac{\#\ word\ types}{\#\ word\ tokens}$			✓
1^{st} person pronouns	*I* *me*			✓
average word log frequency	average log frequency of words in message			✓
n-grams	*I+love* *can't+you+see*	✓	✓	✓
valence	POSITIVE +1.86	✓	✓	
point-wise mutual information	(goodnight, sleepy) = -22.88	✓		
emphasized words	*This is *NOT* okay* *_What_ is that?*	✓		
emoticons	:) ;)	✓	✓	

3 Linguistic features: Going deeper

The basic intuition behind mindprint features is that they capture some knowledge that humans have about a particular mental state. Shallow linguistic features may be useful because they provide a coarse measure of this knowledge. For example, an n-gram such as *the+best* seems to encode some kind of endorsement, which can indicate mental states like confidence or persuasion. However, this notion of endorsement could be encoded more abstractly, as something like *the*+POSITIVE-ADJECTIVE-IN-THE-SUPERLATIVE, and could then be instantiated as a number of expressions: *the best, the brightest, the most fantastic, the most reputable,* and so on. It seems likely that humans use more abstract linguistic features of this kind, since they allow a more compact representation of useful knowledge about language meaning, structure, and polarity. With this in mind, we investigate different abstractions of the words in language text, focusing on automatically derivable semantic, syntactic, and valence features that could be used to create more linguistically abstract n-gram features for mindprints.

3.1 Semantic features

To automatically create more abstract semantic representations for the words in a message, we used the hypernym classifications available through WordNet 3.1 (Fellbaum, 1998; Princeton-University, 2010), where Y is a hypernym of X if X is a kind of Y. For example, *edible-fruit* would be a hypernym of *apple*. We investigated three potential levels of semantic abstraction (each progressively more abstract), shown in Table 2. Only words that had hypernym entries in WordNet were semantically abstracted[2] – all other words in a message were left as is.

Table 2: Example of semantic abstractions of a message, using available hypernyms from WordNet 3.1.

Original Utterance	Level	Semantically abstracted message
Big apples are the best!	1	Big *edible-fruit* are the best!
	2	Big *produce* are the best!
	3	Big *food* are the best!

3.2 Syntactic features

To automatically create more abstract syntactic representations for the words in a message, we used the part-of-speech labels automatically generated by the Stanford Part-of-Speech Tagger (Toutanova, Klein, Manning, & Singer, 2003), which correspond to grammatical categories like PLURAL NOUN and MODAL VERB. We investigated abstracting all words, all

[2]When multiple word senses were available for a word, we selected the most frequent word sense.

content words only (nouns, non-copula and non-auxiliary verbs, adjectives, and adverbs), and all non-content words only, shown in Table 3.[3]

Table 3: Example of syntactic abstractions of a message, using part-of-speech tags from the Stanford Part-of-Speech Tagger.

Original Utterance	Type	Syntactically abstracted message
Big apples are the best!	all	JJ NNS VBP DT JJS!
	content only	JJ NNS are the JJS!
	non-content only	Big apples VBP DT best!

3.3 Valence features

To automatically create more abstract valence representations for the words in a message, which correspond to sentiments like POSITIVE or NEGATIVE, we used the valence ratings from the affective ratings database compiled by Warriner, Kuperman, and Brysbaert (2013), which includes nearly 14,000 words. These ratings ranged between 1 (very negative) and 9 (very positive), and we investigated two options for abstraction (shown in Table 4): replacing all words that had a valence rating with either POSITIVE (if its score was 5 or above) or NEGATIVE (if its score was below 5), and replacing only words whose valence ratings were in the top or bottom third of all valence ratings. This second option effectively abstracted only those words with stronger valences, whether negative or positive.

3.4 The potential features in mindprints

Based on pilot classification results with all 18 combinations of linguistically abstracted features (3 semantic options x 3 syntactic options x 2 valence options), we decided to use the

[3]The labels from the Stanford Part-of-Speech Tagger are as follows, with an example of each tag in parentheses and an indication of whether this tag was considered a content (Co) or non-content (NCo) word: CC=coordinating conjunction (*and*, NCo), CD=cardinal number (*one* penguin, Co), DT=determiner (*the*, NCo), EOS=end of sentence marker (there's a penguin here*!*), EX=existential there (*there*'s a penguin here, NCo), FW=foreign word (*hola*, NCo), IN=preposition or subordinating conjunction (*after*, NCo), JJ=adjective (*cute*, Co), JJR=comparative adjective (*cuter*, Co), JJS=superlative adjective (*cutest*, Co), LS=list item marker (*one, two, three,* . . ., Co), MD=modal (*could*, NCo), NN=singular or mass noun (*penguin, ice*, Co), NNP=proper noun (*Jack*, Co), NNPS=plural proper noun (There are two *Jacks*?, Co), NNS=plural nouns (*penguins*, Co), PDT=predeterminer (*all* the penguins, NCo), POS=possessive ending (penguin*'s*, NCo), PRP=personal pronoun (*me*, NCo), PRP$=possessive pronoun (*my*, NCo), RB=adverb (*easily*, Co), RBR=comparative adverb (*later*, Co), RBS=superlative adverb (*most* easily, Co), RP =particle (look it *up*, NCo), SYM=symbol (this = that), TO=infinitival to (I want *to* go, NCo), UH=interjection (*oh*, NCo), VB=base form of verb (we should *go*, Co), VBD=past tense verb (we *went*, Co), VBG=gerund or present participle (we are *going*, Co), VBN=past participle (we should have *gone*, Co), VBP=non-3^{rd} person singular present tense verb (you *go*, Co), VBZ=3^{rd} singular present tense verb (he *goes*, Co) WDT=wh-determiner (*which* one, NCo), WP=wh-pronoun (*who*, NCo), WP$=possessive wh-pronoun (*whose*, NCo), WRB=wh-adverb (*how*, NCo).

Table 4: Example of valence abstractions of a message, using valence ratings from Warriner et al. (2013)'s affective ratings database.

Original Utterance	Type	Valence abstracted message
Big apples are the best!	all	POSITIVE POSITIVE are the POSITIVE!
	strong only	Big POSITIVE are the POSITIVE!

most successful combination for our main classification task. The most effective semantic abstraction was one level (e.g., *apple* as *edible-fruit* rather than *produce* or *food*), suggesting that too much abstraction loses more nuanced information that humans use to communicate mental states. The most effective syntactic abstraction was content words only (e.g., *Big apples are the best!* becomes JJ NNS *are the* JJS!), which indicates that specific content words are not always so important to the linguistic structures associated with mental states while specific function words contain valuable information. For example, the phrase *the best* is abstracted to *the+JJS* because this construction can indicate confidence or persuasion no matter what superlative adjective is used. In contrast, modal verbs like *can*, *may*, and *should* are linked to different mental states (e.g., confidence (*I can*), formality (*may I*), and persuasion (*you should*)), so abstracting them all to the MD label would destroy this useful distinction. The most effective valence abstraction was the strong valence words only, suggesting that weaker valence information adds noise rather than useful information for detecting mental states. Thus, the linguistic features potentially included in the mindprints we use here include multiple n-gram types: (i) the original n-gram, (ii) the semantically abstracted n-gram, (iii) the syntactically abstracted n-gram, and (iv) the valence abstracted n-gram. The complete list of features is in Table 5, and also includes 43 shallow linguistic features used by Pearl and Steyvers (2013).

4 Language data: Word Sleuth

We investigate the eight mental states spanning intentions, attitudes, and emotions investigated by Pearl and Steyvers (2013): deception (intention), politeness (attitude), rudeness (attitude), embarrassment (emotion), confidence (attitude), disbelief (attitude), formality (attitude), and persuasion (intention). We use language data derived from the Word Sleuth game-with-a-purpose (GWAP) (Pearl & Steyvers, 2013), an online asynchronous game played through a web browser interface (see *wordsleuth.ss.uci.edu*). GWAPs have been used to accumulate information about many things that humans find easy to identify, such as objects in images (von Ahn & Dabbish, 2004; von Ahn, Liu, & Blum, 2006), common sense relationships between concepts (von Ahn, Kedia, & Blum, 2006), beliefs about others' preferences (Hacker & von Ahn, 2009), the musical style of songs (Law & von Ahn, 2009), and mental states communicated through language text (Pearl & Steyvers, 2013). GWAPs leverage human computation (von Ahn, 2006) to produce aggregated results that are more reliable than any particular individual's judgments, a "wisdom of the crowds" effect demonstrated

Table 5: Linguistic features potentially included in mindprints. For all proportion calculations, a smoothing constant (0.5) was added to the raw counts. Note also that lexical diversity values range between 0 and 1, with higher values indicating more diverse usage (each word appears around once). The valence scores were derived from the affective ratings database of Warriner et al. (2013). In addition, all bigrams and trigrams include begin-message (BEGIN) and end-message (END) markers.

Feature type	**Description**	**#**	**Implementation**	**Sample(s)**
basic n-grams	unigrams, bigrams, & trigrams appearing 2+ times in dataset	varies	# of n-gram	*apple* *good+day* *i+love+you*
semantic n-grams	semantically abstracted basic n-grams	varies	# of n-gram	*edible-fruit* *good+time-unit* *i+love+you*
syntactic n-grams	syntactically abstracted basic n-grams	varies	# of n-gram	NN JJ+NN *i*+VBP+*you*
valence n-grams	valence abstracted basic n-grams	varies	# of n-gram	POSITIVE POSITIVE+POSITIVE *i*+POSITIVE+*you*
characters	a,b,c,...z, all digits, all punctuation	28	$\frac{\#\ character(type)}{\#characters}$	*We saw 2 penguins and 3 fish!* $\approx \frac{2}{23}$ digits
punctuation marks	? ! . ; : ,	6	# of mark	*This is the best!* = 1 !
word tokens	number of word tokens	1	# word tokens	*The penguin ate the fish* = 5
word types	number of word types	1	# word types	*The penguin ate the fish* = 4
1^{st} person pronouns	*I, me, my, mine, we, us, our, ours, myself, ourselves*	1	$\frac{\#\ 1^{st}\ person\ pro}{\#\ word\ tokens}$	*We like penguins!* $\approx \frac{1}{3}$
lexical diversity	word type to word token ratio	1	$\frac{\#\ word\ types}{\#\ word\ tokens}$	*The penguin ate the fish* $\approx \frac{4}{5}$
average word length	average characters per word	1	$\frac{\#\ characters}{\#\ word\ tokens}$	*The penguin ate the fish* ≈ 4
average valence score	average word valence	1	$\frac{\sum_{w \in msg} valence(w)}{\#\ words}$	*We saw penguins* $\approx \frac{6.27+6.65}{2}$
average word log frequency	average log frequency of words appearing 2+ times in dataset	1	$\frac{\sum_{w \in msg} log(\frac{\#\ w}{\sum_{d \in ds} \#\ d})}{\#\ word\ tokens\ in\ msg}$	Same as implementation
sentences	number of sentences	1	# sentences	*What did you see? We saw penguins.* = 2
average sentence length	average words per sentence	1	$\frac{\#\ word\ tokens}{\#\ sentences}$	*What did you see? We saw penguins.* $\approx$ 7/2

in many knowledge domains, including human memory (Ditta & Steyvers, 2013), problem solving (Yi, Steyvers, & Lee, 2012), and prediction (Lee, Steyvers, de Young, & Miller, 2012).

In the context of the Word Sleuth GWAP, players can play two roles: the Expressor and the Word Sleuth (see Figures 1 and 2). Expressors are given a specific mental state to express, and provided with a context picture to help guide their message creation. Word Sleuths attempt to determine which mental state was intended in a particular message, and are also shown the context picture that was available to the Expressor during the message's creation. Points are awarded to both the Expressor and the Word Sleuth when a message generated by the Expressor to communicate a particular mental state is accurately interpreted by the Word Sleuth as expressing that mental state. (See Pearl and Steyvers (2013) for more details of Word Sleuth's implementation and game play, including several design features that motivate accurate game play.)

Figure 1: An example of Expressor game play.

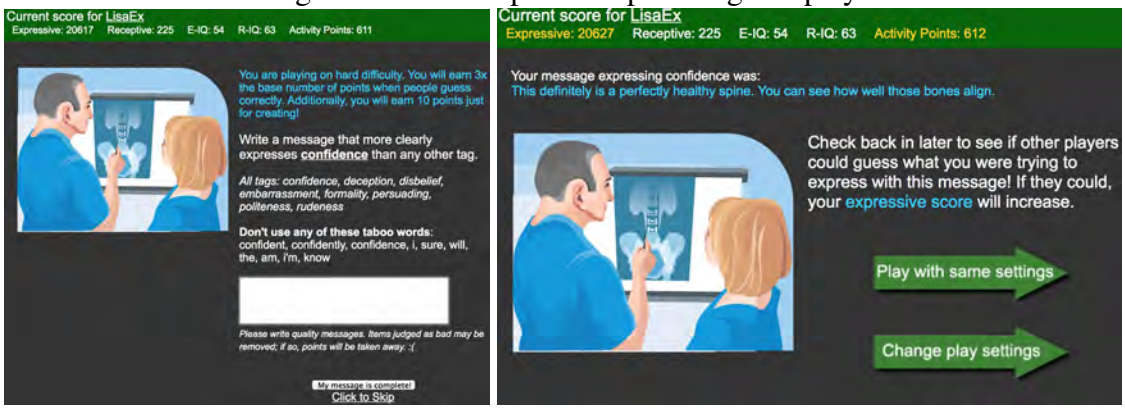
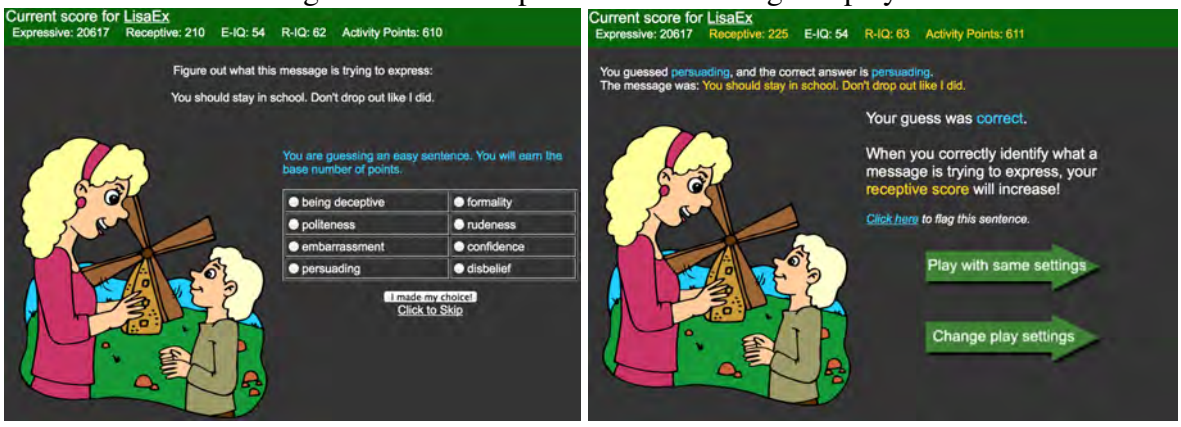


Figure 2: An example of Word Sleuth game play.



There are two main benefits to using a GWAP such as Word Sleuth to generate language data. First, specific mental states can be targeted because players are guided to create messages expressing these mental states. So, instead of researchers sifting through naturally occurring data in an attempt to find language expressing a certain mental state, language is

generated which is known to express that mental state. Relatedly, the "ground truth" is known with respect to which mental state is being expressed in a message, because the Expressor explicitly generated the message to communicate that mental state.

Second, messages that reliably express particular mental states can be identified, based on how accurately a message is interpreted by human players. A message that is consistently perceived to communicate the intended mental state is likely to provide a good source of linguistic cues for that mental state. Thus, we can harness the cumulative interpretations of a collection of humans (the "wisdom of the crowds") to identify messages that are useful for developing mindprint-based techniques to automatically recognize mental states in text.

At the time of writing, the Word Sleuth database contained 4839 messages expressing the eight mental states of interest, and 55577 interpretations of those messages. To identify reliable messages, we selected the subset of messages that had two or more interpretations and greater than 50% interpretation accuracy. Notably, mental states differed on how many reliable messages were generated for them, ranging from 151 to 568. To negate any bias for the most frequent mental state type in the dataset that our classifier would train on, we selected the most accurately perceived 151 messages for each mental state, yielding 1208 messages total with 15514 interpretations. The human accuracy on this dataset was 83%, averaged across messages and participants, with successful mental state interpretation ranging between 70% and 90% (see Table 6), which is significantly better than chance performance of 12.5% (i.e., 1 out of 8). This demonstrates that humans can be very good at transmitting mental states through language text, though still imperfect.

Notably, some mental states are easier than others to express and interpret, as shown in Table 6. The confusion matrix in Table 6 indicates $p(interpreted|generated)$, the probability that a message will be interpreted as expressing a specific mental state (in the columns), given that it has been generated to express that specific mental state (in the rows), averaged over messages and participants. The diagonal probabilities represent how often a message's mental state was correctly interpreted for each mental state, and so indicate mental state transmission accuracy. From this, we can see that rudeness (0.90), confidence (0.89), and disbelief (0.87) are among the easier mental states to express and interpret, while formality (0.70) and deception (0.77) are more difficult. The total number of interpretations for each mental state is shown in the rightmost column.

Table 7 shows sample messages (with the players' own spelling and punctuation), highlighting why some mental states may be easier than others. For example, rude messages tend to use negative valence words, such as "stupid" while confidence can be expressed with markers of certainty like "of course" and disbelief can be expressed with markers of skepticism like "no way." In contrast, formality is often confused with politeness – the confusion matrix in Table 6 indicates that formal messages are often interpreted as polite (0.19) and polite messages are often interpreted as formal (0.09). The underlying issue is that the use of a formal tone is typically a signal of polite discourse (even if the content is negative, such as in a complaint). This causes the linguistic cues used to convey these mental states (e.g., words like "please" in the message in Table 7) to overlap. Notably, it is possible to be polite without being formal, e.g., apologizing for bumping into someone by saying "I'm sorry" is likely to be interpreted as polite, but not formal. Because of this, formality may be inter-

Table 6: Human confusion matrix on the filtered Word Sleuth dataset comprising eight mental states. The rows represent the intended mental state, while the columns represent the interpreted mental state. The bolded diagonal indicates the percentage of correct interpretations for each mental state type. The total number of interpretations for each mental state type is shown in the rightmost column.

	deception	**politeness**	**rudeness**	**embarrassment**	**confidence**	**disbelief**	**formality**	**persuasion**	interpretations
deception	**0.77**	0.02	0.03	0.03	0.02	0.02	0.02	0.08	2316
politeness	0.01	**0.82**	0.01	0.02	0.01	0.01	0.09	0.03	1966
rudeness	0.01	0.01	**0.90**	0.01	0.02	0.02	0.01	0.02	2090
embarrassment	0.02	0.02	0.02	**0.85**	0.01	0.06	0.01	0.01	2341
confidence	0.01	0.02	0.01	0.01	**0.89**	0.01	0.01	0.05	1955
disbelief	0.02	0.02	0.03	0.02	0.02	**0.87**	0.01	0.01	1441
formality	0.01	0.19	0.01	0.01	0.03	0.01	**0.70**	0.04	1723
persuasion	0.04	0.03	0.02	0.04	0.00	0.01	0.02	**0.84**	1682
									15514

preted as a subset of politeness, and so the linguistic cues for formality may be a subset of the cues for politeness. So, to recognize formality, humans must effectively recognize that the message is conveying a particular kind of politeness, rather than politeness in general. This relates to the broader issue of messages conveying multiple mental states, which we return to in section 5.2.1. Since the Word Sleuth players were asked to select the single mental state best represented in a message, different reasonable decisions could be made when multiple mental states were communicated.

Deception is another more difficult mental state for humans to detect, and is often confused with persuasion (0.08). This is also likely due to multiple mental states being communicated in a message, and so having overlapping linguistic cues. In particular, a speaker may attempt deception while trying to persuade the listener of something. This appears in the example message in Table 7: the speaker is trying to persuade the listener that their boss enjoys files being submitted late, even though that is not true. This highlights one way in which deception may be a more complex mental state – it effectively involves the speaker inverting the true semantic content (e.g., their boss does not, in fact, enjoy tardy files) and violating the conversational maxim of Quality (Grice, 1975) that assumes a cooperative conversational partner will be truthful. Thus, deception alters the fundamental content of the message conveyed, while the other mental states do not. For example, a persuasive version of *hate(boss, late files)* does not change the underlying content (e.g., "Our boss really hates late files, so you should make sure to turn yours in immediately."), while a deceptive version would nec-

essarily do so (e.g., "Our boss loves it when you don't give him the files immediately."). Because of this, the linguistic cues for deception are likely more subtle.[4]

Table 7: Sample messages from the filtered Word Sleuth dataset. The top three messages' mental states are correctly interpreted while the bottom two are not.

Intended mental state	**Interpreted mental state**	**Message**
rudeness	rudeness	"Uh, no. Your idea is stupid."
confidence	confidence	"Of course. This case is in the bag."
disbelief	disbelief	"there is no way this perfume is only 20 dollars"
formality	politeness	"Please, calm down. The patient is here and waiting for the results sir."
deception	persuasion	"Our boss loves it when you don't give him the files immediately."

These human confusion data on the filtered dataset are useful in two ways. First, they represent the accuracy levels we would like to achieve (or exceed) with our automatic mindprint-based techniques. Second, they suggest that two of our selected mental states – formality and deception – are likely to be more difficult to automatically classify, since humans sometimes struggle to identify them from language text.

5 Automatically identifying mental states

5.1 The classification task

To set up the classification task, we first determined which machine learning classifier to use. We followed Pearl and Steyvers (2013) and used the Sparse Multinomial Logistic Regression (SMLR) classifier developed by Krishnapuram et al. (2005), since it can use regression analysis to determine the subset of features that are relevant for making its classifications. It assigns these relevant features some weight (depending on how useful they are in actually making a decision), while all irrelevant features are given zero weight. This means that we do not need to identify the useful mindprint linguistic features for each mental state *a priori* – instead, we can allow the SMLR classifier to learn which linguistic features are relevant for detecting each mental state, given all the features available. Thus, while there is a common set of potential mindprint features for mental states, each mental state will have a subset of these actually comprise its mindprint, with that subset identified by the SMLR classifier.

We then determined the baseline of performance to be chance performance when selecting a mental state from one of eight choices (1 of 8 = 0.125), since this is the task the classifier

[4]Sarcasm and irony function similarly by inverting semantic content and violating the conversational maxim of Quality, and so may also involve subtle cues similar to deception cues.

was faced with. Because we eliminated any differences in frequency in the training set, a baseline of selecting the most frequent mental state observed in the training set is equivalent to this baseline.

We used 10-fold classification for training and testing, so that the classifier trained on 90% of the messages and then was tested on the remaining 10%, repeated ten times for each of the folds. Results were then aggregated from all ten iterations, and their average reported.

To use the SMLR classifier, there are two parameters whose values must be set: (i) λ, which determines how strongly the classifier prefers to base its decisions on a smaller subset of features (as opposed to utilizing a larger set), and (ii) r, the number of regression rounds. In our preliminary investigations, we determined that the parameter values yielding the best results were λ=0.0075 and r=10.

5.2 Classifier results

Table 8 shows the mental state identification results from a classifier using the basic linguistic features used by Pearl and Steyvers (2013) (i.e., all linguistic features in Table 5 except for the semantic, syntactic, and valence n-grams), a classifier using the more sophisticated linguistic features described in section 3 (i.e., all linguistic features in Table 5), and human performance on the filtered Word Sleuth dataset. The score shown is the F-score, which is the harmonic mean of precision and recall (F = $2 * \frac{P*R}{P+R}$). Precision is $p(generated|interpreted)$, which is the probability that a message actually was generated to express a particular mental state, given that it was interpreted as expressing that mental state. Recall is $p(interpreted|generated)$ (the accuracy calculation shown in Table 6 above), which is the probability that a message was actually interpreted as expressing a particular mental state, given that it was generated to express that mental state. The F-score provides a single summary statistic that balances precision and recall, as it is possible to have high precision with low recall (e.g., classifying only one message as deception, but doing so correctly) as well as low precision with high recall (e.g., classifying all messages as deception). Thus, to achieve a high F-score, both precision and recall must be high.

We find that a classifier using only basic linguistic features can actually do quite well (average=0.712), similar to the results found by Pearl and Steyvers (2013). However, the classifier that incorporates more sophisticated semantic, syntactic, and valence features improves significantly on this average performance (average=0.756), which suggests that mindprints incorporating these more sophisticated features are indeed helpful. While average performance of the classifier using more sophisticated features does not yet reach human performance (average=0.824), it is much closer. This suggests that the mindprints humans use are likely to contain these more sophisticated linguistic features.

Interestingly, when we examine individual mental states, there is one where the mindprint-based classifiers exceed human performance: formality. This is particularly true of the classifier using more sophisticated linguistic features (Ling-Basic: 0.685, Ling-Soph: 0.721, Humans: 0.656). This suggests that the classifier is better able to distinguish formal cues than humans can, possibly due to the overlap in formal and politeness cues discussed in section 4. This contrasts with the relatively poor performance on deception, the other mental state that

is more difficult for humans. Here, both classifiers struggle more than humans do, though the classifier using more sophisticated features performs better (Ling-Basic: 0.544, Ling-Soph: 0.606, Humans: 0.770).

Table 8: Comparison of the performance of a classifier using basic linguistic features (Ling-Basic), a classifier using more sophisticated linguistic features (Ling-Soph), and humans against the baseline, as measured by F-score.

Mental State	**Baseline**	**Ling-Basic**	**Ling-Soph**	**Humans**
deception	0.125	0.544	0.606	0.770
politeness	0.125	0.702	0.768	0.800
rudeness	0.125	0.638	0.734	0.900
embarrassment	0.125	0.789	0.795	0.874
confidence	0.125	0.794	0.787	0.880
disbelief	0.125	0.755	0.818	0.875
formality	0.125	0.685	0.721	0.656
persuasion	0.125	0.774	0.800	0.835
average	**0.125**	**0.712**	**0.756**	**0.824**

5.2.1 Error patterns

If we wish to bridge the performance gap between automatic identification and human identification, it is useful to know what errors the classifier makes. Moreover, one way to determine if the mindprints used by our classifier are similar to the ones humans use is to see if the classifier errors are similar to human errors. To investigate error patterns, we can examine the confusion matrix for the classifier that uses sophisticated linguistic features. This is shown in Table 9, where $p(interpreted|generated)$ (i.e., the recall score) is calculated for each mental state. The diagonal probabilities represent how often a message's mental state was correctly interpreted for each mental state.

Similar to humans, the linguistically sophisticated classifier is best at identifying confidence (0.84) and disbelief (0.86), rarely confusing them with other mental states. However, unlike humans, this classifier confuses rudeness with politeness (0.07), a curious mistake that Pearl and Steyvers (2013) also found with their original classifier that used shallow linguistic features. Notably, the classifier using more sophisticated features has this confusion less often, which we attribute to its use of valence cues (see Table 11). Nonetheless, this is an indication that our more linguistically sophisticated mindprints still differ from the mindprints humans use to detect rudeness.

For other mental states, the classifier errors are more similar to human errors, suggesting similar mindprints. For example, deception is often interpreted as persuasion (Human: 0.08, Classifier: 0.09). As discussed in section 4, this may not be an unreasonable confusion since many deceptive messages may be attempting to persuade the listener of something. Thus, the

Table 9: Confusion matrix for the classifier using linguistically sophisticated features. The rows represent the intended mental state, while the columns represent the interpreted mental state. The bolded diagonal indicates the percentage of correct predictions for each mental state type.

	deception	**politeness**	**rudeness**	**embarrassment**	**confidence**	**disbelief**	**formality**	**persuasion**
deception	**0.54**	0.01	0.06	0.08	0.12	0.07	0.04	0.09
politeness	0.01	**0.81**	0.05	0.02	0.02	0.02	0.05	0.03
rudeness	0.03	0.07	**0.71**	0.04	0.05	0.05	0.01	0.03
embarrassment	0.07	0.01	0.04	**0.81**	0.02	0.04	0.01	0.01
confidence	0.05	0.03	0.01	0.03	**0.84**	0.01	0.01	0.03
disbelief	0.03	0.03	0.01	0.03	0.01	**0.86**	0.01	0.01
formality	0.04	0.11	0.02	0.03	0.04	0.04	**0.67**	0.05
persuasion	0.02	0.05	0.03	0.01	0.03	0.01	0.04	**0.82**

message really is expressing multiple mental states, but the classifier (and the human Word Sleuth player) is forced to select only one mental state. So, depending on which linguistic features are more salient, the message is interpreted (correctly) as deception or (incorrectly) as persuasion. Notably, because both mental states are likely present in the message, this "mistake" is a reasonable one for the classifier to make. Similarly, the classifier often interprets deception as confidence (0.12), which may also be a legitimate interpretation, as a speaker may attempt to deceive the listener by sounding very confident about the content of the message (e.g., "This is totally and completely safe. I swear.").

Also similar to humans, the classifier interprets formality as politeness (0.11), though notably less so than humans (0.19). This suggests that the classifier may be using mindprint features that humans use, but is able to use them more accurately than humans to distinguish formal messages from polite messages.

5.2.2 Mindprint feature comparison

Since the SMLR classifier learns which subset of the potential linguistic features comprise the mindprint of each mental state, we can examine what these inferred mindprints look like. As Table 10 shows, only a fraction of the available linguistic features were selected for each mental state from the 21,388 potential features. Nonetheless, this small subset was clearly quite useful, since the classifier used those features to achieve the excellent identification performance that it did.

It is also clear from the distribution of these useful mindprint features that the more so-

phisticated linguistic features were useful for every single mental state. The original shallow linguistic features still comprise the majority of mindprint features (ranging between 63% and 72% of the mindprint features), which is likely why Pearl and Steyvers (2013)'s original study (and our replication here) found reasonable identification performance using only those features. Nonetheless, the more sophisticated linguistic features helped bridge the gap between machine and human performance. Also, it turns out that the semantic features were uniformly used more often than the syntactic features, which were uniformly used more often than the valence features for each mental state (Semantic: 12%-18%, Syntactic: 9%-13%, Valence: 3%-6%). This suggests that the semantic features were the most useful, and so more sophisticated semantic features may lead to the largest improvements in automatic mental state identification in the future.

Table 10: The linguistic features in each mental state's mindprint, as inferred by the SMLR classifier. The quantity of features is shown, with the percentage of the total features available (21388) that this quantity represents in parentheses. Also shown is the distribution of mindprint features across the original shallow linguistic features and the deeper semantic, syntactic, and valence features.

Mental State	**Mindprint Features**	**Shallow**	**Semantic**	**Syntactic**	**Valence**
deception	3740 (0.17)	0.72	0.16	0.09	0.03
politeness	2713 (0.13)	0.63	0.18	0.13	0.06
rudeness	3087 (0.14)	0.67	0.16	0.11	0.06
embarrassment	2915 (0.14)	0.72	0.15	0.09	0.04
confidence	2779 (0.13)	0.66	0.17	0.12	0.05
disbelief	2879 (0.13)	0.74	0.12	0.10	0.04
formality	2877 (0.13)	0.68	0.16	0.11	0.05
persuasion	2710 (0.13)	0.68	0.16	0.11	0.05

For each mental state, we can examine the most strongly weighted mindprint features to get a sense of what the linguistic cues for each are (shown in Table 11). In general, the mindprints have at least one kind of linguistically sophisticated feature in the most strongly weighted features, with the exception of deception. Semantic features are found in the strongest features for politeness, rudeness, embarrassment, confidence, and formality; syntactic features are found in the strongest features for embarrassment, confidence, disbelief, formality, and persuasion; valence features are found in the strongest features for politeness, rudeness, embarrassment, confidence, formality, and persuasion. This suggests that not only are these more sophisticated features helpful for constructing the mindprints in general, but they are also highly indicative for most mental states. We also note that all of the strongest mindprint features were positively weighted, and so were associated with the presence of a mental state rather than its absence.

We now discuss some highlights of each mental state's strongest features, particularly in comparison to the mindprints identified by Pearl and Steyvers (2013) (**P&S**). For decep-

Table 11: The most strongly weighted mindprint features for each mental state, including both shallow and deeper (abstracted) features. Semantically abstracted features are *italicized*, syntactically abstracted features are shown with their syntactic label in **bold capitals**, and valence abstracted features are shown in SMALL CAPS. The weight assigned to each feature is shown in parentheses.

Mental State	**Strongest Mindprint Features**
deception	am+an (1.13), nope (0.97), lie (0.97), background (0.81), of+course (0.80), absolutely (0.79), promise (0.79), is+genuine (0.77), um (0.76)
politeness	*desire+your* (2.16), *acknowledgment* (1.82), have+a+good (1.42), lovely (1.30), glad (1.24), want+to+share (1.21), pardon+my (1.18), would+you (1.13), are+a+POSITIVE (1.04), i+POSITIVE+your (1.01), BEGIN+would (1.00)
rudeness	*dislike* (2.01), screw+you (1.54), you're+NEGATIVE (1.47), jerk (1.43), NEGATIVE+NEGATIVE+END (1.42), *misfit* (1.32), idiot (1.23), ew (1.18), *waste* (1.18), trash (1.12), is+NEGATIVE+END (1.13)
embarrassment	shame (2.27), *clothcovering* (1.71), 1^{st} person pronouns (1.27), ashamed (1.24), oops (1.20), clumsy (1.11), oh+**NN** (1.11), i+POSITIVE+NEGATIVE (1.01), a+fool (0.84), he+NEGATIVE (0.78)
confidence	*control* (1.67), *emotion* (1.60), the+**JJS** (1.23), BEGIN+im (1.12), i+can+**VB** (1.12), am+POSITIVE (1.09), can+**VB**+this (0.99), i'm+POSITIVE (0.98), awesome (0.96), mvp (0.93)
disbelief	unreal (1.21), cannot (1.08), BEGIN+really (1.05), BEGIN+wow (0.87), no+way (0.86), this+**RB** (0.79), she+**RB** (0.77), he+**RB** (0.76)
formality	may+i (1.60), pardon+me (1.51), BEGIN+good (1.40), *aristocrat* (1.40), introduce (1.37), how+**VBP** (1.22), order (1.19), allow (1.17), may+i+POSITIVE (1.11)
persuasion	try+it (1.34), BEGIN+you+should (1.28), BEGIN+come (1.06), you're gonna (0.86), should+POSITIVE (0.84), c'mon (0.83), you+should+**RB** (0.80), if+you+**VBP** (0.73), you+should+**VB** (0.72)

tion, we find the presence of verbs indicating intention ("promise") and indicators of uncertainty ("um") as P&S did. However, we also find indicators of certainty ("of+course", "absolutely"), which are more subtle cues that correlate with the persuading intention and confident attitude that can accompany deceptive messages. So, while the semantic reversal that deception entails is difficult to identify solely from linguistic features, these certainty indicators can signal intentions and attitudes associated with deception. For politeness, we find that specific positive valence words ("lovely", "glad") and phrases involving modals ("BEGIN+would", "would+you") are strongly correlated, similar to P&S. In addition, we see that phrases involving words with positive valence ("are+a+POSITIVE", "i+POSITIVE+your") are highly indicative of politeness, presumably because one way to be polite is to compliment the listener.

For rudeness, we find that specific negative valence words ("idiot", "trash") are highly correlated, as P&S found. We additionally find that phrases that contain semantically abstracted forms of specific negative valence words (e.g., *misfit* for "dork" and "jerk") and negative valence words in general are highly indicative ("you're+NEGATIVE", "is+NEGATIVE+END"). This is intuitively satisfying, as a simple way to be rude is to insult the listener, which can be captured by negative valence words. For embarrassment, we find that phrases indicating the appearance of an accident ("ashamed", "shame", "oops", "clumsy") are correlated, as P&S found. In addition, we find that first person pronouns are highly indicative, perhaps because a common cause of embarrassment is something the speaker is responsible for. We also find that negative valence words are highly indicative ("he+NEGATIVE"), which is likely because causes of embarrassment are typically negative. For confidence, we also find that first person pronouns are highly indicative, as P&S found. In addition, many of the more sophisticated linguistic features are highly correlated: semantically abstracted features (e.g., *control* for "conquer" and "handle"), syntactically abstracted features involving the modal "can" ("can+**VB**+this", "i+can+**VB**"), and positive valence words associated with first person pronoun markers ("am+POSITIVE", "i'm+POSITIVE''). For disbelief, we find that indicators of surprise ("BEGIN+wow", "unreal"), some of which involve negation ("no+way", "cannot"), are highly correlated, as P&S found. We additionally find that syntactically abstracted features are highly indicative, particularly those involving adverb abstractions of "really" or "actually" ("he+**RB**", "she+**RB**", "this+**RB**"), which express surprise. For formality, we find some fixed formal expressions ("pardon+me", "may+i") are correlated, as P&S found. In addition, we find that the semantically abstracted formal title (*aristocrat* for "highness" and "prince") is highly indicative, suggesting this more general representation of a formal title was very useful. For persuasion, we find coercive expressions ("you're gonna", "c'mon", "try+it", "BEGIN+you+should") are correlated, as P&S did. In addition, syntactically abstracted features allow instantiations of the phrase "you should..." ("you+should+**RB**", "you+should+**VB**") to rise to the top of the persuasion mindprint features. Positive valence words following the coercive modal "should" are also highly indicative ("should+POSITIVE"), perhaps because expressing something in a positive light can make it sound more attractive to the listener.

From this, it is apparent that the more linguistically sophisticated features are not only generally helpful for constructing mindprints but usually among the strongest indicators of a mental state. Thus, they are capturing some of the deeper knowledge that humans use to detect mental states in language. Still, because automatic mental state identification has not reached human levels for most mental states, it is likely that other linguistic features can be used to augment mindprints. Based on the results here, it is likely that more sophisticated semantic features could capture some of the missing abstract representations that humans use. For example, potentially useful semantic abstractions include *uncertainty* (deception), *certainty* (deception), *intention* (deception), *accident* (embarrassment), *surprise* (disbelief), *formal expression* (formality), and *coercion* (persuasion).

Some potential tools exist for automatically identifying these semantic classes. The Linguistic Inquiry and Word Count database (Pennebaker, Booth, & Francis, 2007) was developed to examine emotional, cognitive, structural, and process components present in lan-

guage. It includes explicit lists of words that cover *uncertainty* (cognitive processes: tentative) and *certainty* (cognitive processes: certainty). Similarly, WordNet-Affect (Strapparava & Valitutti, 2004) was developed to aid in emotion identification research, and includes a WordNet class that explicitly lists words related to *surprise*.

Still, there are some potentially useful semantic classes that do not currently have explicit lists of words available (*intention, accident, formal expression, coercion*). For these, we may be able to use a machine learning technique called topic modeling (Griffiths & Steyvers, 2004) to automatically identify words corresponding to these semantic classes. "Topics" in this approach are probability distributions over keywords that relate to a cohesive "concept", broadly construed. In particular, a topic may be something we typically think of as a concept, such as *fictional villains*, or instead something that represents a set of expressions that share a stylistic component, such as *casual expressions*. These topics, and the keywords that comprise them, are identified in an unsupervised fashion from a collection of documents. Without any additional information beyond the documents themselves, topic models can use the words contained in the documents to identify both the topics expressed and which topic each word, sentence, or subsection of the document most likely belongs to. Given a topic model trained over a large enough collection of documents, we may find that a topic model can spontaneously create the list of words associated with some of the semantic classes of interest. For example, Pearl and Steyvers (2012) trained a topic model on a collection of blog texts, and discovered a *casual expressions* topic containing expressions such as "oh", "lol", "yeah", and "gonna". Since such words would not appear in messages expressing a formal attitude, this is a semantic class that should be strongly negatively correlated with formality. It may thus be possible to automatically identify useful semantic classes for mental states using this technique, and perhaps even the specific semantic classes identified above that are likely to indicate the presence of a particular mental state.

In addition, it may be useful to take more inspiration from the way humans process information when we consider linguistic features. For example, the semantic features considered here always abstracted a word one level up, using WordNet hypernyms (e.g., *Dalmatian* would be abstracted to *dog*). Interestingly, *dog* is what Rosch (1978) terms a "basic level" category of semantic representation, and is a level frequently used by humans when referring to objects. Given this, a more nuanced semantic abstraction might collapse only those words that are more detailed than this basic level, while leaving basic level words alone (e.g., *dog* would not be collapsed to *canine* as it was here). While it is non-obvious how to automatically identify the basic level for words from WordNet currently, it may be possible to draw from the semantic categorization literature to discover a way to do so.

Also, while our focus in this study was on more general-purpose linguistically sophisticated features, it is likely that augmenting the mindprints proposed here with features incorporating domain-specific knowledge about a mental state will improve identification results (e.g., hedges for politeness, affect lexicon features for embarrassment). This would combine the insights from previous work on targeted mental state identification (e.g., politeness, emotions) with the results here that highlight the utility of deeper purely linguistic features.

5.3 Mindprint complexity and generalizability: Future work

There are other natural extensions to the mindprint-based approach we have presented here. First, we can recognize the complexity of mental states that messages can communicate and identify multiple mental states simultaneously in text. Our current findings indicate that messages can be perceived as expressing multiple mental states (e.g., Table 6: deception and persuasion), which suggests that the mindprints of these mental states can (and often do) occur in the same message. So, we may wish to allow both our mindprint-based classifiers and our human Word Sleuth players to identify multiple states when that happens. Given this additional information about the mental states expressed in messages, we may then be able to extract more nuanced mindprints for a particular mental state (e.g., separating out the deceptive and persuasive components of a message that reflects both those mental states).

Second, we may wish to test how general the mindprints are that we identified here, and how generalizable this mindprint-based approach is to other naturalistic texts. To do this, we would want a corpus of naturalistic linguistic data that is annotated with the mental states expressed by those data. It has typically been difficult to find this type of reliable annotation for naturalistic corpora of significant size, which was one of the motivations for the creation of the Word Sleuth game (see Pearl and Steyvers (2013)). Nonetheless, a good existing resource to start with would be the Livejournal blog corpus of Mishne (2005), where each entry has been tagged with the mood the author was in when writing that entry. Mishne (2005) and Keshtkar and Inkpen (2009) used that mood tag as an indicator of the mental state the entry conveyed (though it is likely each entry conveyed additional mental states as well). If we assume that the mood an entry is tagged with is definitely communicated in the entry's text (e.g., an entry tagged as *embarrassed* expresses embarrassment), we can investigate two questions: First, how well do the mindprints we have identified here work on the Livejournal data? For example, do embarrassed entries use features of the embarrassment mindprint we learned from the Word Sleuth data? Second, how well does the mindprint-based approach work for other mental states encoded by the Livejournal moods? For example, can we automatically identify mindprints for *happy*, *contemplative*, and *sleepy* moods using our approach? If our approach is truly a general approach for identifying mental states in text, it should be successful for other texts and other mental states.

6 Conclusion

We have investigated the utility of automatically constructed mindprints for identifying a variety of mental states spanning intentions, attitudes, and emotions in text. Importantly, these mindprints use more sophisticated linguistic features than previous studies have used, and capture some of the more abstract knowledge humans may use when detecting the linguistic signature of a mental state. By using more linguistically sophisticated features, our mindprint-based technique achieves near-human level performance when identifying most mental states from text and exceeds human-level performance for one mental state. This indicates that mindprints incorporating deeper linguistic knowledge are a valuable tool for

intelligent systems that are conversing with humans, as such systems can more easily identify the subtle information about mental states that humans convey in language.

7 Acknowledgements

We are very grateful to Marge McShane, three anonymous reviewers, and the members of the Computation of Language laboratory at UC Irvine for valuable comments and suggestions. We are also indebted to Kristine Lu and Athenia Barouni for their comments on the mental state identification literature, and all the Word Sleuth players who generated the data used in this study. This research was supported by a UC Irvine Summer Undergraduate Research Program Fellowship to Igii Enverga.

References

Anand, P., King, J., Boyd-Graber, J., Wagner, E., Martell, C., Oard, D., & Resnik, P. (2011). Believe me – We can do this! Annotating persuasive acts in blog text. In *Proceedings of the AAAI Workshop on Computational Models of Natural Argument.* San Francisco, CA: AAAI.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge, MA: Cambridge University Press.

Chaffar, S., & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. *Lecture Notes in Computer Science*, *6657*, 62–67.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of ACL.* Sofia, Bulgaria: ACL.

Ditta, A., & Steyvers, M. (2013). Collaborative memory in a serial combination procedure. *Memory*, *21*, 668–674.

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics. 3: Speech acts.* (pp. 41–58). New York: Academic Press.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*, 5228-5235.

Hacker, S., & von Ahn, L. (2009). Matchin: Eliciting user preferences with an online game. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 1207–1216). Boston, MA: Association for Computing Machinery.

Hardisty, E., Boyd-Graber, J., & Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of Empirical Methods in Natural Language Processing* (pp. 284–292). Boston, MA: ACL-EMNLP.

Keshtkar, F., & Inkpen, D. (2009). Using sentiment orientation features for mood classification in blogs. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineerings* (pp. 24–29). Dalian, China: IEEE.

Krishnapuram, B., Figueiredo, M., Carin, L., & Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 957–968.

Kruger, J., Epley, N., Parker, J., & Ng, Z.-W. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology*, *89*(6), 925–936.

Law, E., & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 1197–1206). Boston, MA: Association for Computing Machinery.

Lee, M., Steyvers, M., de Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, *4*, 151–163.

Lin, W., Wilson, T., Wiebe, J., & Hauptmann, A. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the Conference on Natural Language Learning (CoNLL).* New York: CoNLL.

McPartland, J., & Klin, A. (2006). Asperger's syndrome. *Adolescent Medicine Clinics*, *17*(3), 771–88.

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics.* Singapore: ACL.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of SIGIR 2005.* Salvador, Brazil: ACM.

Mohammed, S. (2012). Portable features for classifying emotional text. In *Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 587–591). Montreal, Canada: NAACL-HLY.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)* (p. 806-814). Beijing China: COLING.

Pearl, L., & Steyvers, M. (2010). Identifying emotions, intentions, & attitudes in text using a game with a purpose. In *Proceedings of NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles, CA: NAACL.

Pearl, L., & Steyvers, M. (2012). Detecting authorship deception: A supervised machine learning approach using author writeprints. *Literary and Linguistic Computings*, *27*(2), 183–196.

Pearl, L., & Steyvers, M. (2013). "C'mon – You should read this": Automatic identification of tone from language text. *International Journal of Computational Linguistics*, *4*(1), 12–30.

Pennebaker, J., Booth, W., & Francis, M. (2007). *Linguistic inquiry and word count: Liwc*. Austin, TX: LIWC.net.

Princeton-University. (2010). *About WordNet.* http://wordnet.princeton.edu.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorizatio.* Hillsdale: Lawrence Erlbaum Associates.

Rubin, V. L., & Conroy, N. J. (2011). Challenges in automated deception detection in

computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing* (pp. 1556–1560). Fortaleza, Brazil: ACM.

Strapparava, C., & Valitutti, A. (2004). WordNetAffect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 1083–1086). Lisbon, Portugal: LREC.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003* (pp. 252–259). Edmonton, Canada: HLT-NAACL.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer Magazine*, *June*, 96–98.

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 219–326). Vienna, Austria: Association for Computing Machinery.

von Ahn, L., Kedia, M., & Blum, M. (2006). Verbosity: A game for collecting common-sense facts. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 75–78). New York, NY: Association for Computing Machinery.

von Ahn, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (pp. 55–64). New York, NY: Association for Computing Machinery.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Yi, S., Steyvers, M., & Lee, M. (2012). The wisdom of crowds in combinatorial problems. *Cognitive Science*, *36*(3), 452–470.