Behav Res (2018) 50:344–361
DOI 10.3758/s13428-017-0875-9

Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis

Distributed dictionary representation

Justin Garten[1] · Joe Hoover[1] · Kate M. Johnson[1] · Reihane Boghrati[1] · Carol Iskiwitch[1] · Morteza Dehghani[1]

Published online: 31 March 2017

Abstract Theory-driven text analysis has made extensive use of psychological concept dictionaries, leading to a wide range of important results. These dictionaries have generally been applied through word count methods which have proven to be both simple and effective. In this paper, we introduce Distributed Dictionary Representations (DDR), a method that applies psychological dictionaries using semantic similarity rather than word counts. This allows for the measurement of the similarity between dictionaries and spans of text ranging from complete documents to individual words. We show how DDR enables dictionary authors to place greater emphasis on construct validity without sacrificing linguistic coverage. We further demonstrate the benefits of DDR on two real-world tasks and finally conduct an extensive study of the interaction between dictionary size and task performance. These studies allow us to examine how DDR and word count methods complement one another as tools for applying concept dictionaries and where each is best applied. Finally, we provide references to tools and resources to make this method both available and accessible to a broad psychological audience.

Keywords Methodological innovation · Text analysis · Semantic representation · Dictionary-based text analysis

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0875-9) contains supplementary material, which is available to authorized users.

✉ Morteza Dehghani
mdehghan@usc.edu

[1] Computational Social Science Laboratory, University of Southern California, Los Angeles, CA 90089, USA

Language and communication play a central role in psychological research both as direct objects of study and as windows into underlying psychological processes. In order to automate analysis of large quantities of text-based communication, psychological researchers have primarily captured psychological phenomena by developing and applying domain dictionaries (Stone et al., 1968; Pennebaker et al., 2001), lists of words which are considered indicative of a particular latent factor. These dictionaries have generally been applied using word-count methods that involve counting the frequency of dictionary words in samples of text. This intuitive approach has provided a simple method of applying domain knowledge to large sources of data. This has been successfully applied to topics ranging from sentiment analysis (Tausczik & Pennebaker, 2010) to group differences in moral concerns (Graham et al., 2009) to the evaluation of depression in clinical patients (Ramirez-Esparza et al., 2008).

This work has also led to insights which have fed back into both linguistics and computer science. One notable discovery has been the importance of closed class terms to understanding psychological properties from language (Pennebaker, 2011). A number of word classes such as determiners, pronouns, and conjunctions and subclasses such as modal verbs are considered to be closed since they are relatively fixed with words rarely added or removed. Given the Zipfian distribution of language (Powers, 1998), these small sets of common words compose around 60 % of many English texts. Preferring to focus on content words, many computational approaches dismissed these as "stopwords" (Wilbur & Sirotkin, 1992) which could be safely ignored. However, psychological applications of dictionaries and word counts showed these to be essential to understanding a range of phenomena including emotional state (Pennebaker, 1997),

authorship identification (Boyd & Pennebaker, 2015), and social hierarchies (Kacewicz et al., 2013).

While word count is an ideal method for applying psychological dictionaries composed of closed-class terms, many dictionaries which have attempted to codify psychological concepts (such as positive and negative emotions, terms associated with depression, morally loaded terms, etc) are composed of open-class terms. These terms, taken from classes such as nouns, adjectives, and verbs are much larger than closed classes (even if the individual terms are less frequently used) and provide a range of unique challenges.

First, even a well-developed open-class dictionary will struggle to capture a concept in all possible contexts. No researcher can be familiar with all possible sociolects (Louwerse, 2004), that is, all dialects associated with combinations of age groups, ethnic groups, socioeconomic classes, gender cohorts, regional clusters, etc. There is no such thing as domain independence when it comes to language. This can have pernicious effects, as measures will be most effective when applied to groups similar to the researchers and their immediate cohort (Henrich et al., 2010). Medin et al. (2010) referred to this as "home-field disadvantage." While it's possible to bring in representatives of particular groups of interest, this vastly increases the complexity of dictionary generation and is still limited to groups which the researchers are aware of.

Second, the contextual dependence of language means that a simple list of open-class words can only cover narrow strips of a concept. Even in the simple domain of product reviews, words can have opposite senses within a single category. For cameras, "long" is positive in the sense of having a "long battery life" while negative when referring to a "long focusing time" (Liu, 2010; Iliev et al., 2015). A cold beer is good while a cold therapist is probably best avoided.

Third, the dynamic and generative aspects of language mean that even a theoretically universal dictionary would not remain so for long. While this is less of an issue in some contexts such as fixed sets of historical texts (Smith et al., 2000), it is particularly salient in modern online contexts such as social media. Terms rapidly appear, disappear, and change meanings. Linguistic resources which expect consistent usage, "correct" grammar, or even recognizable spelling are of limited use for text coming from Facebook or Twitter (Kouloumpis et al., 2011). While lexical drift might not pose a major short-term threat to dictionary validity, it is likely that as the time between dictionary construction and application increases, the chance of error increases. Dictionaries can be updated, but this requires yet more resource investment.

All of this combines to make dictionary development both resource intensive and challenging to do well. While the most representative words for a category might be easy to recall, low-frequency words can easily escape even the most diligent expert. Variations in colloquial lexicons across social and cultural groups make implicit bias a constant threat.

Systems such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) have helped both by providing a validated set of high-quality dictionaries and decreasing the cost of dictionary based text analysis by providing pattern matching tools. A pattern like "ador*" (from the LIWC affect dictionary) would automatically capture words like "adore", "adoration", and "adoringly". However, because this pattern matching relies on morphological similarity, semantically unrelated terms can also be caught (e.g., "adornment" in the previous example). Given the risk of such unwanted terms, knowing how and when to use these patterns required additional expertise[1] which further complicates dictionary authorship.

One possible answer is to consider semantic similarity rather than morphological. This is a the approach we explore here, applying psychological dictionaries by measuring their similarity to words or segments of text rather than asking whether or not words are explicitly in the dictionary.

Of course, this raises the question of how to measure semantic similarity. To this end, we make use of distributed representations, where words are represented as points in a low-dimensional space (generally 10-1000 dimensions). Such spaces offer a simple way to determine similarity between words in terms of distance in the space. For example, two words which are highly similar to one another, such as "doctor" and "physician", would be near one another in the space.

While often treated as a recent development, distributed representations have been explored for decades. Geometric spaces were first used to represent semantic structure early as the 1950's in psychology (Osgood et al., 1957) and continued to develop in the information retrieval community through the 1970's (Salton et al., 1975). One of the more popular approaches for psychological applications, Latent Semantic Analysis (LSA) (Deerwester et al., 1990), emerged by combining work from the information retrieval community with psychological research on word meaning and language learning (Landauer and Dumais, 1997).

More recently, cognitive psychologists developed Parallel Distributed Processing (PDP) (Rumelhart et al. 1988), where distributed representations approached their current form. Work on PDP not only demonstrated how these representations might be used in complex tasks but also showed how they could be learned from and serve as the natural inputs and outputs of neural networks.

[1] Such as awareness of relative word frequencies in the target domains to know when spurious wildcard matches can be safely ignored.

Critically, these multi-dimensional spaces proved to be relatively easy to generate using large bodies of unlabeled text. The primary approach has been based on the distributional hypothesis, a formalization of J.R. Firth's aphorism, "you shall know a word by the company it keeps" (Firth, 1957). Effectively, this says that if two words occur in similar contexts, they're likely to be more similar to one another. So, if we saw the sentences "the cat ran across the room" and "the dog ran across the room", we could infer that cats and dogs likely had some things in common. While a single instance might not tell us much, over billions of words and contexts, we are able to build highly detailed representations.

Modern methods of generating distributed representations (Collobert & Weston, 2008; Mikolov et al., 2013; Pennington et al., 2014) prove to be both more efficient and produce representations that show surprising semantic regularities.[2] For example, the nearest neighbors of terms tended to be highly meaningful. Given the word "frog", the nearest terms in one representation were "frogs", "toad", and then the names of particular species of frogs ("litoria", "leptodactylidae", "rana") (Pennington et al., 2014).

Further, it was found that representations trained in this way did not just encode attributional similarity (a measure of synonymy between words) but also relational similarity (a measure of analogical similarity between two pairs of words) (Medin et al., 1990; Turney, 2006). Simple linear transformations encoded meaningful semantic, syntactic, and structural relationships. This famously allowed for the solution of simple analogical reasoning problems such as "man is to woman as king is to (queen)" (Mikolov et al., 2013).

For psychological research, these techniques offer a unique opportunity. In particular, the structural regularities observed in distributed representations provide a route past some of the challenges around applying dictionaries of open class words.

In the current paper, we introduce the Distributed Dictionary Representation (DDR) method and explore its application to two domains, one learning to predict the sentiment of movie reviews and another attempting to determine the moral loading of posts on Twitter. Finally, we carry out an extensive evaluation of how a dictionary's size and structure impacts its effectiveness when applied through DDR.

Distributed dictionary representations

We demonstrate a novel method of combining psychological dictionary methods and distributed representations which indicates that these two methods are not only compatible, but that combining the two adds to the flexibility of both and opens new avenues for exploration. Our method, which we term Distributed Dictionary Representation (DDR), averages the representations of the words in a dictionary and uses that average to represent a given concept as a point in the semantic space. We can use this representation to provide a continuous measure for how similar other words are to a given concept.

One advantage of this method is to improve the ability to apply dictionaries to small pieces of text (down to individual words). This is critical as more and more social scientific text analysis makes use of social media posts (Mitchell et al., 2013; Kern et al., 2014; Eichstaedt et al. 2015; Dehghani et al., 2016) which are often no more than a few words long. At that length, it is unlikely that *any* words from an open-class dictionary will be present to be counted. Prior social media research has noted precisely this difficulty (Gunn & Lester, 2015) with the common solution being to aggregate multiple short posts into larger documents (Tumasjan et al., 2010). The disadvantage is that we lose post-level granularity and the ability to track changes over time, critical in a number of areas such as clinical psychology.

DDR also has a number of benefits for dictionary development. Since the purpose of the dictionary is now to identify the core of a concept rather than identifying every possible word which might be associated with that concept, it is possible to produce a dictionary with a small list of the most salient words. This makes it easier for researchers to generate new dictionaries and apply them to explore theoretical concepts where the resources may not have been previously available for large-scale text analysis. By making use of distributional semantic similarity, researchers can focus on concept validity rather than dealing with linguistic issues.

Method details

The goal of the DDR method[3] is to take a list of words characteristic of a category (often referred to as a concept dictionary in social scientific research) and use those words to generate a continuous measure of similarity between that concept and any other piece of text. The key factor that separates this from the standard application of a dictionary using word count methods is the combination with a pre-trained distributed representation. This representation can either be trained on a wide-coverage corpus of text (web pages, news articles, etc) or on a domain specific corpus (social media posts, interview transcripts, etc).

[2] Although recent work has suggested that these regularities also existed in LSA (Levy & Goldberg, 2014), this was obscured due to non-linearity.

[3] Available as open-source software at https://github.com/USC-CSSL/DDR/.

DDR creates a concept representation by finding the vector representation of each of the words in the dictionary and averaging them. This is similar to the averaging method for generating sentence or document representations from word embeddings (Landauer & Dumais, 1997; Foltz et al., 1998; Mitchell & Lapata, 2008). Using this concept representation, other words, phrases, or documents can be compared to determine their similarity to the category.

For example, imagine that a psychologist wished to study happiness and, for simplicity, assume that their dictionary consisted only of the words "happy" and "joy". Given a distributed representation, DDR would find the vector representations of these two words and average them to produce a concept representation based on this dictionary.

Formally, we consider a non-empty dictionary D of m words $\{w_1, w_2, \ldots, w_m\}$ and a pre-trained n dimensional distributed representation R. R can be treated as a map defined over the words in its vocabulary V such that, for each word in its vocabulary, R maps that word to an n-dimensional real-valued vector:

$$R(w) = [d_1, d_2, \cdots, d_n], \forall w \in V$$

So, to take a word from our previous example, R would map "joy" to an n-dimensional vector corresponding to the word's location in the distributed representation.

The next step is to generate the representation of the concept dictionary C_R in the chosen distributed representation R. We first find which words in the dictionary are present in the vocabulary of the distributed representation, taking the intersection of the dictionary D and the vocabulary V:

$$D_R = D \cap V$$

Finally, we add the representations of the words in this intersection together and normalize this value to generate a concept representation compatible with the word representations in the distributed representation:

$$C_R = \frac{\sum_{\forall w \in D_R} R(w)}{\| \sum_{\forall w \in D_R} R(w) \|}$$

With this category representation C_R, we can now calculate its similarity to any word w in the vocabulary V. We make use of cosine similarity,[4] a measure which defines similarity in terms of the angle θ between the vectors.[5] Similarity is maximized when $\cos\theta = 1$ meaning that $\theta = 0$ (i.e. the vectors are parallel). A value of -1 signals that that $\theta = \pi$ (i.e. the two are pointing in opposite directions and so are maximally dissimilar). Cosine similarity returns a value between 1 (maximum similarity) and -1 (maximum dissimilarity) and can be calculated by:

$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$$

where $\|x\|$ is the length of vector x:

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

This formula allows us to make use of a computational shortcut where, by pre-normalizing all vectors in the space,[6] the previous calculation reduces to $\cos\theta = x \cdot y$.

We apply a similar method for determining the similarity of a document or phrase to a dictionary. As a first step, a document summary vector, T_R is generated using the same word-averaging method described for generating the concept representation. That is, we find the representation for each word in the document in the distributed representation, add each of those together, and normalize the resulting vector.[7]

We can make use of this document representation in the same way we previously made use of word vectors. In particular, we can find the similarity of the document representation T_R to the concept representation C_R with the same distance metric we discussed previously to return a final similarity score ranging between 1 and -1:

$$D(C_R, T_R) = C_R \cdot T_R$$

Given this, we can combine a dictionary with a distributed representation to produce a measure of category similarity in terms of that representation. Just as applying a dictionary through word counts yields a single scalar value (in that case, the normalized count of the words in a document found in the dictionary divided by the total number of words in the document), DDR produces a single scalar value representing the distributional similarity of the dictionary to the document.

Understanding DDR

It's useful at this point to look at an examples to see how DDR works and what information is being captured by the underlying concept of semantic similarity. In Study 2, we make use of the Moral Foundations Dictionary (Graham et al., 2009), an operationalization of Moral Foundations Theory (Haidt et al., 2009) which posits five key moral domains (see Study 2 for details).

[4] While this simple symmetric notion of similarity has been shown to be inadequate by previous research (Tversky, 1977; Medin et al., 1990), it is still useful. Prior work has shown that the local structure of nearest neighbors for terms are highly semantically relevant (Jones & Mewhort, 2007; Mikolov et al., 2013) even if they don't capture the full psycholinguistic notion of concept similarity.

[5] For normalized vectors, this is order-equivalent to Euclidean distance

[6] This is usually done in a single pre-processing pass for the entire distributed representation.

[7] While this method of generating document representations ignores word order and the syntactic structure of the document, this sort of "bag of words" representation is sufficient for many domains.

Fig. 1 Nearest neighbors of the authority MFD domain

Fig. 2 Nearest neighbors of the word "authority"

One of the dictionaries we consider is designed to specify the authority virtue domain and is composed of the words "authority", "obey", "respect", and "tradition". By creating a representation of this dictionary in DDR, we can directly examine the nearest neighbors of this category.[8] When considered using distributed representations trained on the full text of Wikipedia, the top 10 nearest words we find are: "obedience", "deference", "regard", "adherence", "uphold", "govern", "obeyed", "affirm", "dignity", and "respecting". In Fig. 1 we show an expanded view of the nearest neighbors of this category.

We can compare this to the results for the individual words in the category. For example, the nearest neighbors of the word "authority" alone are "jurisdiction", "government", "authorities", "responsibility", and "commission" (see Fig. 2). The nearest neighbors of "respect" are: "regard", "deference", "respecting", "fairness", and "disrespect" (see Fig. 3). Each of these words has a slightly different focus than the average of the dictionary. As we would hope, the combination of words helps to clarify the concept we wish to explore.

We can see this as well on a larger, more extensively validated dictionary, the positive emotions category from LIWC. Here, the nearest neighbors of the dictionary are: "endearing", "earnestness", "heartfelt", "captivating", "youthful", "exuberant", "likable", "amiable", "carefree", and "alluring" (see Fig. 4). Once again, the dictionary representation seems to be capturing the kinds of terms we would hope to catch with this method.

Beyond the direct application of existing dictionaries, distributed representations can also assist in the process of dictionary creation. Given a set of words, finding the most distributionally similar terms can help to spur further development. Given a few words which a dictionary author believes to be highly relevant, looking for distributionally similar terms can not only help to suggest new words but also indicate when a given term has alternative senses which may obscure the intended category.

With these examples, it becomes easier to understand some of the differences between applying a dictionary through word counts and DDR. In particular, we can consider this in terms of the differences in how a dictionary can go wrong. Let's say we had a dictionary of 100 terms associated with depression. We could add millions of obscure scientific terms to this dictionary and, while it would destroy its face validity, it would not affect its performance when applied through word counting to a set of standard psychological interviews (assuming that none of the patients were discussing particular gene pathways and such). However, those terms would completely change the performance of this dictionary when applied through DDR since those terms would completely shift the representation of the dictionary.

On the flip side, adding a single high frequency term to our original dictionary (such as "the" or "a") would have minimal impact on its performance with DDR. That one term would not shift the averaged dictionary representation. However, through word count, the frequency of such a common term would have a large impact on the results of that dictionary.

This points to the different strengths and weaknesses of these two methods when applying psychological dictionaries. Word count is far more sensitive to high frequency terms, being part of the reason for our suggestion to prefer this approach when dealing with closed class terms. DDR is more sensitive to the semantic relation of the terms in the dictionary, leading to our suggestion to prefer it with open class terms. Further, we believe it is part of the reason why smaller dictionaries may work better with DDR as it is easier to maintain semantic coherence in these cases.

In subsequent experiments, we both explore this comparison and particularly focus on the impacts of different dictionaries and distributed representations on the overall effectiveness of DDR.

[8]The available code makes it easy to explore this for other words and dictionaries.

Fig. 3 Nearest neighbors of the word "respect"

Study 1: sentiment analysis

One of the key uses of psychological text analysis is in the inference of the intents and attitudes which underlie statements. Language is more than just a means of expressing a collection of facts (in spite of logical positivism's (Carnap, 1959) best efforts). The critical questions often revolve less around *what* was said than *why* it was said.

A number of the factors behind the production of a piece of text have been explored, but one of the most studied is the simplified notion of sentiment analysis (Liu, 2010)—is the writer or speaker positive or negative about the topic under discussion? Extensive work has been done in terms of products ranging from movies to presidential candidates (Godbole et al., 2007; Pang & Lee, 2008).

Much of this has been driven by ease of combining the text of reviews with discrete annotations such as star ratings for movie reviews (Pang & Lee, 2005). This has allowed for the simplified creation of large labeled datasets, ideal for the application of unsupervised learning methods. Approaches have continued to evolve, both in terms of problem formulation (Chen et al., 2015) and as the full weight of modern machine learning techniques have been brought to bear (Socher et al., 2013; Tai et al., 2015; Li et al., 2015).

In this study we focus on predicting the polarity of movie reviews. Movie reviews provide an interesting domain as they have been shown to be among the more difficult areas for sentiment analysis (Turney, 2002). This is due to a number of issues including the blend of writing about the movie and writing about *elements* of the movie (a movie about a failed heist is not necessarily a failure), the tendency to offer separate appraisals for elements of the movie (e.g. "a wonderful performance that was more than the writing deserved"), and the range of genres (a "hilarious" comedy is good while "hilarious" might be an insult if applied to a movie attempting solemnity).

Our aim is not to compete with state of the art classification methods, but rather to examine how and when distributed representations can allow us to extract a clearer signal when applying dictionaries to text. We compare DDR against word count methods, using both to generate equivalent features which we evaluate in terms of performance on a downstream polarity classification task. Within DDR, we evaluate different dictionary sizes and compare the effects of using different distributed representations.

Method

We make use of a dataset of 2000 movie reviews (Pang & Lee, 2005) introduced to explore machine learning techniques for sentiment analysis. This widely-used dataset (Mcauliffe & Blei, 2008; Socher et al., 2011; Li & Liu, 2012) was originally obtained from the Internet Movie Database (IMDb) archive.[9] Reviews were automatically sorted as positive or negative based on star ratings. To limit the impact of individual authors, there was a limit of 20 reviews from any single writer. The full dataset of all labeled reviews can be downloaded from the Cornell archive.[10]

We compare two general approaches for applying dictionaries–word count and DDR. For a given dictionary and piece of text, word counting returns a value between 0 and 1, representing the normalized count of dictionary words in the document. DDR returns a value between -1 and 1, representing the distributional similarity of the dictionary to the document.

In order to keep the evaluation as consistent as possible, we standardize several factors across the experiments. All classification is done using logistic regression (Hosmer &

[9] http://www.imdb.com/reviews/

[10] http://www.cs.cornell.edu/people/pabo/-movie-review-data/

Fig. 4 Nearest neighbors of the LIWC positive emotions dictionary

Lemeshow, 2004). While not the highest performing classification method available, it has the virtue of model simplicity while maintaining sufficient power to handle issues such as differing means of independent variable values (critical for this dataset).

All evaluations are done on the full set of 2000 documents with 10-fold cross validation. We evaluated results in terms of F1 score (Powers, 2011), which is calculated as the harmonic mean of precision and recall. Precision (or positive predictive value) evaluates the ratio of true positives to total predicted positives of a classifier while recall (or sensitivity) measures the ratio of correctly predicted positives to the total size of the class. By considering the harmonic mean of these two values, F1 balances these factors.

The first method we evaluate is a direct application of the LIWC (Pennebaker et al., 2001; Tausczik & Pennebaker, 2010) word count method and dictionaries to this dataset. In particular, we count instances of words in the positive emotions (containing words such as: "love", "nice", and "sweet") and negative emotions categories (containing words such as "hate", "ugly", and "annoyed") for each of the documents. Based on prior evaluations of psychological dictionaries, we chose to use the LIWC (Tausczik & Pennebaker, 2010) positive and negative categories over other dictionaries such as PANAS-X (Watson & Clark, 1999). Not only is LIWC is widely used, these dictionaries have been shown to be more effective for sentiment analysis on this dataset (Frimer & Brandt, 2015).

However, while prior studies made use of the positive and negative emotion LIWC dictionaries, we wanted to confirm that this was in fact a valid choice. As such, we performed word counts for all LIWC 2007 dictionaries. We then calculated the information gain (Lindley, 1956; Box & Hill, 1967; Fedorov, 1972), a means of measuring the capacity of a variable to reduce uncertainty, for the results from each of the dictionaries. The positive and negative emotion dictionaries had gains of 0.0270 and 0.0170 respectively (while these values are low, in this case we care primarily about the relative informativeness of the dictionaries). The only other two LIWC categories in this range were negation with an information gain of 0.0214 and discrepancy with a gain of 0.0177. Given that prior work had focused on positive and negative emotion dictionaries, we chose to focus on these categories. Negation and discrepancy seemed to be picking up the tendency of certain reviews to equivocate (e.g. 'good acting but...'). While this would be an interesting phenomenon for future exploration, we felt it to be beyond the scope of the present paper.

To generate features for use in classification we first ran the basic LIWC word count (including morphological matching) to get a total count of the words in the document and the words for the selected dictionaries. Given this, we found the percentage of the document composed of positive and negative words and used these values as features for a logistic regression model.

With DDR, we tested several combinations of dictionaries and representations. We made use of three representations, one publicly available set[11] trained on approximately 100 billion words from Google News articles,[12] one trained on the full text of the English Wikipedia,[13] approximately 2.9 billion words in total, and one trained on approximately 11 million words from movie reviews[14] beyond those in our test set.

All distributed representations were trained using Word2Vec (Mikolov et al., 2013).[15] Given the different training sets, each distributed representation had a different vocabulary size. The Google News representations had a vocabulary of approximately 3 million words, the Wikipedia representations had a vocabulary of approximately 2 million words, and the IMDb representations had a vocabulary size of approximately 45,000 words.

While the sizes of these spaces were very different, we felt that this corresponded to a common research situation. In many cases, researchers have access to large quantities of open domain text or even pre-trained distributed representations while having access to a much smaller amount of data in their focal domain. Thus, the choice of whether to make use of general purpose representations trained on more data or more focused representations trained on less data is salient to many real-world tasks.

The LIWC dictionaries make extensive use of pattern matching (e.g. providing root patterns rather than complete

[11] Available at https://code.google.com/p/word2vec/.

[12] http://news.google.com/

[13] https://dumps.wikimedia.org/

[14] Available at http://ai.stanford.edu/~amaas/data/sentiment/.

[15] Making use of the skip-gram with negative sampling model with the following parameters: window 10, negative 25, hs 0, sample 1e-4, iter 15.

words such as "ador*" to capture "adore", "adoration", etc). Since DDR makes use of the representations of individual words, we first expanded the LIWC patterns against a separate corpus of movie reviews. Originally, the positive emotions dictionary contained 405 words and patterns and the negative emotions dictionary contained 500. Positive emotions expanded to 1286 full words and negative emotions expanded to 1718 words.

For use with DDR, we compared these expanded LIWC positive and negative emotions dictionaries with a small set of representative words chosen to explore task-specific dictionary creation. For the task-specific dictionaries (which we refer to as seed dictionaries), we chose 4 words characteristic of positive and negative reviews. For the positive, we chose ["great", "loved", "amazing", and "incredible"]. For the negative we used ["hated", "horrible", "crap", and "awful"].

Using DDR, we combined each of the two dictionary pairs (expanded LIWC and seed) with each of the three distributed representations. We used each of these to generate two features for the movie reviews in our test set, a measure of the similarity of the review to the positive and negative categories. This yielded six sets of features, each of which we could directly compare against the word count method.

Results

All results (see Table 1) were based on 10-fold cross validation to minimize the effects of overfitting (especially with the relatively small test set). Reported values are averaged over 10 iterations of 10 fold cross-validation. Significance was estimated via 10,000 iteration permutation testing with paired-sample t-tests.

Using the LIWC dictionaries with the standard word count method yielded an F1 score of 0.658, in line with prior work considering this method (Kahn et al., 2007).

Table 1 Results for Study 1: performance on the 2000 document movie sentiment corpus

Model	Precision	Sensitivity	F1
Full LIWC Dictionary - Word count	0.657	0.659	0.658_a
Full LIWC - Wikipedia embeddings	0.659	0.649	0.654_a
Full LIWC - IMDb embeddings	0.695	0.682	0.689_b
Full LIWC - Google News embeddings	0.715	0.699	0.707_c
Seed LIWC - Wikipedia embeddings	0.665	0.654	0.660_a
Seed LIWC - IMDb embeddings	**0.764**	**0.762**	$\mathbf{0.763}_d$
Seed LIWC - Google News embeddings	0.745	0.723	0.734_e

Subscript letters indicate significant difference from other score, $p < 0.0001$, calculated using random permutation tests

Numbers in bold represent the largest values in that column

For the DDR tests, when combined with Wikipedia-derived representations, neither the LIWC nor the seed dictionaries showed a significant improvement over the word count results. However, the results were different when making use of representations trained on Google News and IMDb. When making use of the full dictionary with IMDb and Google News-derived vectors, the DDR F1 scores improved to 0.697 and 0.707, respectively. Notably, when using the seed dictionary, the combination with IMDb and Google News representations further increased this improvement, yielding F1 scores of 0.763 and 0.734. To determine whether F1 scores were significantly different between feature sets, random permutation tests with 10,000 iterations were conducted. Specifically, for each feature set, 1,000 10-fold models were estimated and the F1 score from each fold was extracted, yielding 10,000 F1 scores for each method. Random permutation testing was then performed to test the null hypothesis that sampled F1 scores for each method were drawn from the same population distribution. This approach is recommended because random permutation tests are robust to violations of normality and F1 scores are known to not be normally distributed, which violates the assumptions of the dependent-samples t-test (Smucker et al., 2007; Menke & Martinez, 2004).

Discussion

This study demonstrates how, with the proper choice of distributed representations, DDR can provide benefits both for classic, extensively validated dictionaries and for a potentially new style of dictionary generation. Although, it should be seen as augmenting rather than replacing existing best practices. While a great deal of work in the computational realm focuses on raw performance results, for social scientific research, model interpretability remains a key factor when trying to draw explanations of the underlying concepts. These results suggest that we can combine some of the benefits of both theory-driven and bottom-up methods.

In terms of dictionary generation, these results point to the potential to apply dictionary methods to novel areas of social scientific research. While the ability to develop a large, psychologically and linguistically validated dictionary remains out of the reach of most teams, it is far more feasible to find a small set of keywords corresponding to a given concept of interest. In many studies, this already takes place in the early phases of theory validation.

In conjunction with DDR, these small sets of core words are sufficient to allow for large scale text analysis, either in support of concept validation or in an application domain as demonstrated here.

While the task-specific seed dictionary performed better in this case, this result should not be overly generalized. A set of words chosen as characteristic of a given domain

should generally outperform a more broad-coverage dictionary. What is noteworthy here is less the performance difference than the fact that simplified dictionary generation worked at all. With word count methods, small dictionaries generally have too few hits to generate a viable signal on most documents (too few words from the dictionary are present in any given document). The ability of DDR to generate a clean signal with smaller dictionaries has the potential to open up a variety of tasks and theoretical constructs to text analysis.

The statistically significant improvement in the performance of the LIWC dictionaries when applied with DDR compared to word count demonstrates that these benefits can be realized with established dictionaries. This is a critical test as it would have been just as easy to imagine this type of expansion as diluting the overall value of a large dictionary rather than enhancing it.

The variations in DDR performance based on the choice of distributed representation was one of the most intriguing aspects of this study. For example, DDR with representations trained on Wikipedia performed notably worse for both seed and full LIWC dictionaries. We believe that the reason for this is that the task was focused on determining sentiment while Wikipedia explicitly rejects the inclusion of personal sentiment or opinions in articles.[16] As such, the very information critical for this task may not have been present in the distributed representation. Comparatively, the Google News training and IMDb corpora contained a larger representation of sentiment-relevant contexts. While we caution against overinterpretation of the fact that the IMDb distributed representation produced the best overall performance in this test, we do find it notable that a domain-specific distributed representation trained on a fraction of the data performed comparably to much larger representations.

It is also worth noting, that while the Google news embeddings outperformed the IMDb embeddings when full-dictionary features were used, the opposite was the case for the seed-dictionary models. While we explore this issue further in Study 3, we have not yet determined a simple rule for predicting the precise interaction of a dictionary and representation with DDR.

In sum, this study provides evidence for two key findings. First, it shows that using small sets of domain-central terms to construct concept dictionaries is a viable technique when combined with DDR. Second, regardless of the generation technique, when paired with appropriate representations, DDR is able to improve the performance of dictionaries on application tasks. This offers both a new route for dictionary generation as well as a new approach to applying dictionaries to text analysis tasks.

[16]Wikipedia editing instructions specifically require articles to be written from a "neutral point of view" https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

Study 2: the morality of Twitter

In this study, we continue comparing the application of dictionaries through DDR and word count methods in a more difficult context—the identification of moral rhetoric in social media posts. This task has a number of features that make it useful as a follow-up to the previous study. First, rather than considering the binary task of positive/negative sentiment, this task is both multi-class (we consider 10 moral categories and an additional non-moral class) and multi-label in that a single post can include multiple moral dimensions (except for the exclusive non-moral category). These make the identification of moral rhetoric difficult even for human raters. Second, social media posts are far shorter and noisier than movie reviews, complicating the task of extracting a clean signal. Third, the domain of moral rhetoric lacks external annotation (such as star ratings for movie reviews) around which to construct a labeled dataset. As such, this case corresponds more closely to the common situation in psychological analysis where validation and application of an underlying theory interact with one another.

To represent moral sentiment, we make use of Moral Foundations Theory (Graham et al., 2009), which classifies moral concerns into five domains: Care/harm (sensitivity to the suffering of others), Fairness/cheating, (reciprocal social interactions and the motivations to be fair and just when working together), Loyalty/betrayal (promoting in-group cooperation, sacrifice, and trust), Authority/subversion (endorsing social hierarchy), and Purity/degradation (promoting cleanliness of the soul over hedonism) (Haidt et al., 2009).

These categories have been operationalized using the Moral Foundations Dictionary (MFD; Graham et al. 2009), a collection of terms representative of the positive (virtue) and negative (vice) aspects of each foundation, which yields a total of 10 moral sentiment dimensions.

We consider moral rhetoric in the context of Twitter posts collected during period surrounding Hurricane Sandy, specifically looking at posts calling for donations. This analysis faces a set of challenges raised by the brevity of these messages which have a famously fixed limit of 140 characters (Tumasjan et al., 2010).

The short length means that, compared to the article-length movie reviews analyzed in the previous study, there is very little text to work with. Given this restriction, comments often have a greater dependency on shared contexts and cultural nuance. References are terse and unexplained,

jokes are brief and allusive, and short-hand is the norm. Further, the casual nature of the medium means that unconventional spelling and grammar (both intentional and un-) is rampant, complicating the task of natural language processing. In short, though Twitter has been the focus of a great deal of psychological social media research, it remains an exceptionally difficult domain of analysis.

As in the previous study, we test the quality of the signal extracted with a given dictionary by comparing features derived from word count and DDR on the task of identifying moral rhetoric.

Method

To create this data set, we sampled 3000 Tweets from a set of 7 million posts related to Hurricane Sandy between October 23, 2012 and November 5, 2012. The raw set was filtered to exclude retweets and those lacking location information, then limited to Tweets discussing donation. Three trained coders each coded 2000[17] of these Tweets on 11 dimensions—the five Moral Foundations broken down into virtues and vices and a "non-moral" dimension, which was used to indicate that a given Tweet did not contain moral content. Excluding the non-moral dimension, all dimensions were permitted to overlap so that a given Tweet could be coded as containing moral rhetoric relevant to multiple moral concerns. Coders were trained over multiple sessions by first being introduced to the overall MFT framework with subsequent sessions detailing the domains and covering potential ambiguities. They were not specifically trained on the MFD.

Given the low base rate of expressions of moral sentiment, we pre-selected Tweets based on their nearness to the distributed semantic spaces representing each moral domain. Specifically, for a given moral dimension, we selected the 250 Tweets that loaded highest on that dimension, yielding a sample of 2500 Tweets. To ensure that non-moral Tweets would also be represented in the sample, an additional 500 Tweets were randomly selected from a subset of Tweets that, across all dimensions, had moral loadings that fell within one standard deviation of 0.

We used the three sets of independently coded data to compare the information value of different feature sets. First, we tested word counts based on the MFD dictionaries. We directly counted the words in the MFD categories and represented each Tweet as a 10-dimensional feature vector based on the normalized counts for each category.

Second, we applied DDR as in Study 1. We made use of the Google News and Wikipedia distributed representations discussed in Study 1 but replaced the IMDb trained representation with another publicly available set trained using Twitter data as the domain-specific comparison set. This set was trained using GloVe (Pennington et al., 2014) on 2 billion tweets with a resulting vocabulary size of approximately 1.2 million words.[18] For all three representations, we created separate concept representations for each of the 10 MFD categories. Then, for each tweet, we calculated the distance between the tweet and the 10 concept representations to yield features for use in classification.

Third, we continued our exploration of dictionary size and distributed representation expansions. In consultation with the original MFD authors, we selected four words most representative of each of the MFD categories. These seed dictionaries were then applied as in the previous step where, for each of the 10 MFD categories, we found the concept representation of the seed dictionary and used the distance between these concept representations and each tweet to generate 10 features for use in classification. The chosen seed words are listed in full in Table 2.

Table 2 Seed words selected for each of the MFD categories

MFD category	Seed words
Authority virtue	authority obey respect tradition
Authority vice	subversion disobey disrespect chaos
Care virtue	kindness compassion nurture empathy
Care vice	suffer cruel hurt harm
Fairness virtue	loyal solidarity patriot fidelity
Fairness vice	cheat fraud unfair injustice
Loyalty virtue	fairness equality justice rights
Loyalty vice	betray treason disloyal traitor
Sanctity virtue	purity sanctity sacred wholesome
Sanctity vice	impurity depravity degradation unnatural

As in the previous study all classification was done using logistic regression with 10-fold cross-validation. Because the rate of positive codes within each coded dimension was unbalanced (e.g. for a given dimension, many more Tweets were coded as not containing rhetoric relevant to that dimension than were coded as containing relevant rhetoric), positive cases were upsampled by selecting cases with replacement from the lower-frequency class. Comparisons were once again made in terms of F1 score.

Results

As we made use of human-annotated data as the gold standard for this task, evaluation of inter annotator agreement was key. Agreement was measured using Prevalence and Bias adjusted Kappa (PABAK) (Byrt et al., 1993; Sim & Wright, 2005), an extension of Cohen's Kappa robust to unbalanced data sets. PABAK, which can be evaluated

[17] Each with an overlap of 1000 Tweets.

[18] Available at http://nlp.stanford.edu/projects/glove/.

using the same rough guidelines as Kappa, was reasonably high for all dimensions (for the moral dimensions averaged across coder pairs, M = 0.81, SD = 0.07).

All classifiers were evaluated on precision, sensitivity, and F1 score for each of the 10 MFD categories. Significance of F1 differences across methods was calculated using permutation testing with 10,000 iterations (Table 3).

In all of the cases we examined for this study, features derived from the application of dictionaries through distributed representations using DDR significantly outperformed features derived from word count methods. This was true across all three of the distributed representations we considered.

Of the three distributed representations explored, the best results for both the full MFD and the seed subset were generated through the combination with the Google News vectors. For these, using the full MFD yielded an F1 of 0.485 while the seed set yielded a significantly higher F1 of 0.496.

Features generated from the Wikipedia distributed representations yielded the lowest overall performance of the DDR tests with F1 from the full MFD of 0.405 and seed MFD of 0.411. The Twitter derived distributed representations were slightly better than the Wikipedia results with the full MFD features yielding an F1 of 0.421, which was significantly better than either of the Wikipedia tests and significantly worse than either of the two Google News derived feature sets. The Twitter seed MFD features were not statistically different from the Wikipedia full MFD features with an F1 of 0.415.

While seed features significantly outperformed those derived from the full MFD when concept representations were generated in terms of the Google News and Wikipedia models, features derived from the seed dictionary were significantly worse when generated using the Twitter distributed representations.

Table 3 Results for Study 2: method performance averaged across coders and dimensions

Model	M Precision	M Sensitivity	M F1
Full MFD - word count	0.181	0.457	0.275_a
Full MFD - Google News	0.363	0.837	0.485_b
Full MFD - Wikipedia	0.294	0.758	0.405_c
Full MFD - Twitter	0.312	0.764	0.421_d
Seed MFD - Google News	**0.372**	**0.840**	$\mathbf{0.496}_e$
Seed MFD - Wikipedia	0.302	0.755	0.411_f
Seed MFD - Twitter	0.305	0.763	0.415_f

Subscript letters indicate significant difference from other score, $p < 0.0001$, calculated using random permutation tests

Numbers in bold represent the largest values in that column

Discussion

This study supports the results of Study 1 that applying open class dictionaries through DDR is able to extract a clear signal. The combination of greater conceptual and task complexity suggests that these results may generalize across a range of domains and contexts.

These results were particularly interesting in terms of the applicability to short-text such as social media posts. In this case, the relatively low performance of features derived from word count methods is unsurprising. Many of the posts included no words from any of the MFD dictionaries meaning the classifier could do no better than chance and often seemed to overfit the limited signal available. When dealing with short text fragments, the ability of DDR to generate a finer-grained measure of similarity is valuable.

This has implications for more than just social media. Even when longer texts are available, DDR allows researchers to consider components of the documents. By taking measurements at the paragraph, sentence, or even subsentential levels, DDR allows for consideration of how usage shifts over the course of discourse or topical shifts.

Given the structure of the task in terms of the detection of 10 highly related concepts, it was unclear if DDR would provide useful information or blur the dimensions. However, the results here suggest that even with these more finely structured and related categories (compared to positive/negative sentiment in the previous task), DDR is able to improve the detection of signal. This further suggests that there is in fact semantic separation among the various MFD categories.

In terms of the comparison between the full MFD and a small subset taken from each category, this study reinforces our findings on the applicability of much shorter lists of words than found in traditional psychological dictionaries. While the seed dictionaries outperformed the full MFD in two of three cases (when combined with the Google News and Wikipedia representations), the important factor is not the strong claim that dictionary authors *should* use smaller word lists but rather than more modest claim that they *can* make use of smaller lists.

Further, this highlights the value of DDR for dictionaries that are still in the developmental process. Not all dictionaries are as extensively developed or well-validated as those found in LIWC and this study suggests the utility of DDR in precisely this case. Ideally, this would allow other researchers to explore concepts prior to extensive validation and refinement of an underlying dictionary. Given only a handful of the most salient words, DDR simplifies the process of applying and refining those concepts.

Study 3: dictionary selection

The previous two studies demonstrated the utility of the combination of dictionaries and distributed representations. In both studies DDR was better able to measure the similarity between a dictionary and a piece of text than classic word count methods. However, while we observed differences in the effectiveness of particular combinations of dictionaries and distributed representations for particular tasks, the reasons for those differences were unclear.

In particular, while we considered both large, validated dictionaries and smaller, "seed" dictionaries, it was unclear how much we could generalize from the observed results. In both studies, we observed that seed dictionaries generally performed as well or better than the full dictionaries (with the exception of the combination with the Twitter representations in Study 2). However, with so few examples, it was unclear whether this was due to the particular examples we considered and what factors would affect this.

To consider these questions about DDR, we repeated the structure of Study 1 with 5.1 million unique dictionary pairs (10.2 million total dictionaries). To generate these, we sampled from the LIWC positive and negative emotions dictionaries to generate dictionaries ranging from 2 to 900 words. We evaluated each pair of generated positive and negative dictionaries on the same movie review sentiment task of Study 1 allowing us to compare how dictionary size interacts with task performance.

With this data, we were further able to explore how the structure of the dictionaries affects the applicability to a downstream task. Since DDR works by averaging the representations of words in the dictionary to generate a concept representation, a natural starting point is to consider the similarity of the selected words. In particular, we evaluated how the clustering of dictionary words in the distributed representation affected resulting classification performance.

Method

In this study, we make use of the same framework as in Study 1, using DDR-derived features to predict the polarity of movie reviews (Pang & Lee, 2005). We generate dictionary pairs by sampling from the LIWC positive and negative emotion dictionaries. For distributed representations, we compare the two top performing distributed representations from Study 1—the IMDB-trained representations and the Google News representations.

As one of our key questions related to the ideal size of a dictionary, we considered a range of dictionary sizes. Given that the intersection of dictionary and distributed representation vocabulary varied, we separately sampled for each of the two distributed representations.

Dictionary pairs were generated by sampling from the intersection of the expanded LIWC positive and negative emotion dictionaries with the distributed representations. For IMDb representations, this yielded intersections of 888 and 1203 words for the positive and negative dictionaries. For the Google News distributions the intersections were 988 and 1383 words. We sampled words without replacement from each of these two sets to create dictionary pairs of length $2, 3, \ldots, 10, 20, \ldots, 100, 200, \ldots, 900$ for a total of 26 separate sizes (the IMDb representations did not include the 900 case as the intersection wasn't large enough). For each size, we generated 100,000 dictionary pairs for each of the two distributed representations yielding 5.1 million pairs total.

For each of these dictionary pairs, we repeated the experiment of Study 1: first generating concept representations for each of the dictionaries using DDR, then finding the distance of documents to those concept representations, and finally using those distances as features to train a classifier for sentiment polarity. For each of the resulting feature sets, we evaluated performance using 10-fold cross validation, reporting averaged F1. Within each of the sample sizes, we compared several measures of overall performance: the mean over all samples, the mean of samples two standard deviations above the mean, and the best overall sample.

We also used these results to explore whether variations in dictionary structure affected resulting classifier performance. In particular, we wanted to see whether dictionaries which were more semantically clustered (that is, where the words of the dictionaries were nearer to one another in the semantic space) led to better resulting performance when combined with DDR.

Towards this, we evaluated the semantic coherence of each generated dictionary. For each pair of words in the dictionary, we calculate their cosine similarity and average those together for all such pairs. This requires $\binom{n}{2}$ calculations for a dictionary with n words. As this is $O(n^2)$, it is tractable for any reasonably sized dictionary.

We calculated this all-pairs similarity for each of the 10.2 million dictionaries (from the 5.1 million dictionary pairs) and used the resulting measures for each dictionary pairs as features to predict the resulting F1 scores by fitting with a linear regression model. We did this for each sample size, calculating an R^2 for each size.

Results

We started by looking at the mean performance over all samples at each size in order to compare how dictionary size affected performance. The results can be seen in Fig. 5 (complete results can be seen in tables Tables 1 and 2 in the

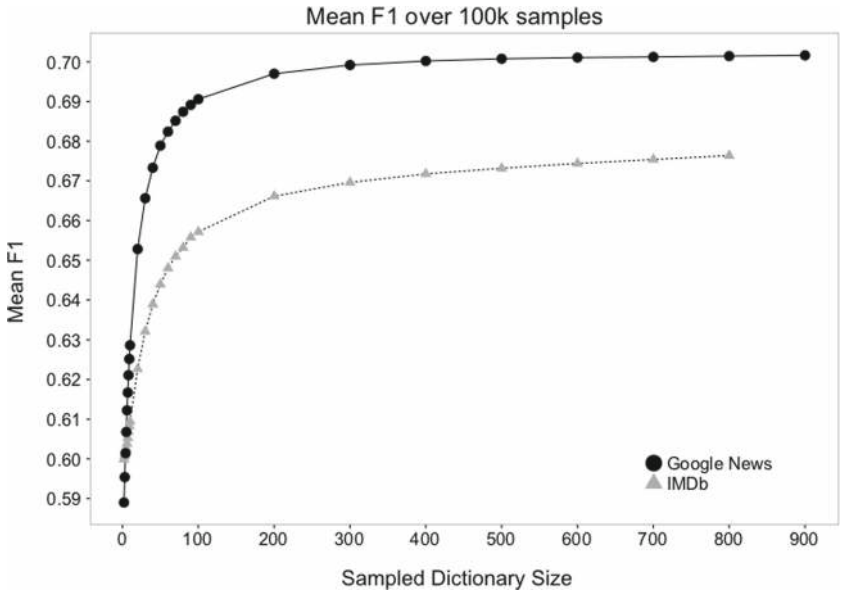

Fig. 5 Mean F1 of dictionary sizes

Supplementary Material). The mean performance asymptotically approaches the performance of the full dictionary in terms of both of the two distributed representations.

For these results, the large number of samples made traditional significance tests irrelevant (with 100,000 samples for each sample group, almost any difference is significant) so we focused instead on measures of effect sizes. In particular, we made use of Cohen's *d* (Cohen, 1988) to compare the predictive power of the size of the dictionary pairs. Table 3 in the Supplementary Materials provides a complete pairwise comparison of the set of samples for each dictionary size. While dictionaries with similar numbers of words showed only marginal differences (mean value of 0.156 for adjacent sizes), larger gaps yielded very large effect sizes (the d between samples of size 2 and 100 was 2.950). The large numbers of samples kept the variance small (< 0.0005 for all reported values).

Given that we are randomly sampling from each of the two dictionaries, this result makes intuitive sense. As we increase the sample size, those samples increasingly closely approximate the complete set of words. At smaller sizes, we're more likely to capture sets of words whose DDR representation diverges significantly from the overall concept representation.

However, our concern is not with randomly sampled dictionaries, the aim is to determine how size affects ideal performance. As such, we looked specifically at sampled dictionary pairs which performed better than 2 standard deviations above the mean for each sample size. As seen in Fig. 6 we see a very different pattern here (raw results can be found in Tables 4 and 5 in the Supplementary Material). Overall performance rises quickly to dictionaries of 30 words, subsequently falling towards the performance of the full dictionaries.

It is interesting to note the difference in performance between making use of representations trained on Google News and IMDb text. In this case, combining dictionaries with IMDb representations produces better results with smaller-sized dictionaries while doing worse with larger dictionaries. The performance difference at the large scale fits with the previous results where Google News representations performed better when using the full LIWC dictionaries (see Table 1 in Study 1). As the sample sizes increase, we would expect the results to increasingly closely approximate those results. However, it is unclear why the opposite was observed for smaller dictionaries.

We next looked directly at the best performing dictionaries for each sample size. Although ordinarily this would be an outlier, given the large number of samples at each size, it is reasonable to consider these best cases. Although there is noise in this data, as seen in Fig. 7 (and Table 6 in the Supplementary Materials), the overall pattern matches that seen with the results for the +2 standard deviation case. Once again, performance increases up to dictionaries of length 30 (40 for the Google News vectors) and subsequently declines. We also see the same pattern where IMDb representations outperform at smaller sizes while declining more rapidly as dictionary size increases.

Finally, given this large stock of dictionary pairs and performance results, we considered the interaction of semantic similarity and resulting classifier performance. While we expected that semantic coherence would be positively corr

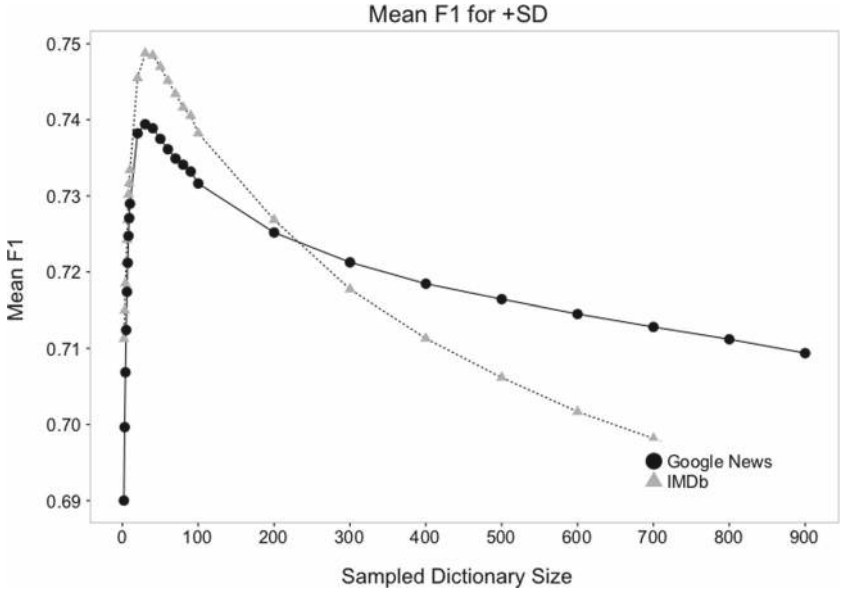

Fig. 6 Mean F1 of dictionaries performing 2 standard deviations above the mean

elated with dictionary performance, we found an almost complete lack of correlation. The maximum observed coefficient of determination was 0.209 with a mean of 0.017 across all sample sizes.

Discussion

This study confirms that, with DDR, smaller dictionaries can outperform larger ones. In particular, this confirms that the performance of the seed dictionaries in the previous two studies were not merely due to chance or selection effects. At the same time, some of the particular findings of this study raise a number of questions for future research.

One issue we would caution against is overgeneralizing from the particular numbers in this case. While we found 30 word dictionaries to produce the best overall results on the downstream classification, there is nothing to suggest that this particular number would generalize to other applications, tasks, or dictionary types.

However, the overall structure of the results strongly support the findings of the first two studies that smaller dictionaries are sufficient when applied through DDR. In

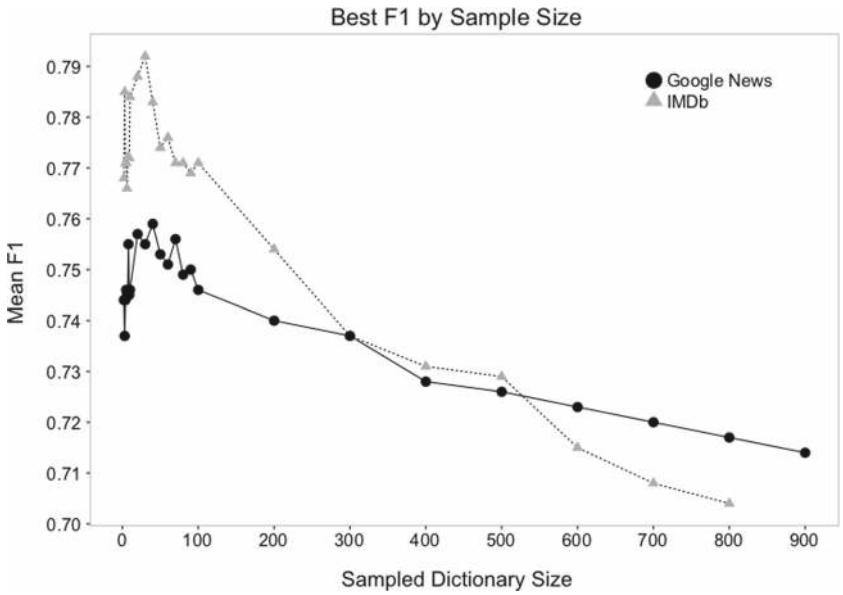

Fig. 7 Best performing sample at each size

particular, we note the case of the best performing samples at each size where the difference in resulting performance was small for dictionaries smaller than 100 words. When combined with IMDb representations, the resulting performance ranged between F1 scores of 0.766 and 0.792 while for Google News those scores ranged between .737 and 0.759.

This suggests that, for a well designed dictionary, length is not a critical factor. Dictionary authors should feel free to incorporate as many or as few words are necessary to get at the desired theoretical construct without undue focus on size (unless sampling at random which doesn't seem to be a common technique for dictionary generation). This is a style of dictionary which is only possible when applied through DDR as word count methods depend on linguistic coverage.

We were somewhat surprised that, for these examples, dictionaries of size 30 produced the best overall results. This was larger than we'd expected given our prior work. However, as noted, we don't believe that this particular number can be applied without validation to other combinations of domain and dictionaries. As before, given the small differences in performance, we feel that concept validity should be the guide rather than any other optimality criteria.

Also surprising was the lack of correlation between semantic coherence and resulting performance. While our prediction had been that more semantically similar dictionaries would perform better, we found no correlation. In hindsight, this makes sense given the experiment and DDR's structure. As we sampled from coherent categories, the particular sampling choices are more capturing different facets of a single concept rather than separate concepts. In terms of DDR, what matters is the generated concept representation, not the particular words used to generate it. Two dictionaries might have completely different words and structures yet generate similar concept representations. This points to a number of possibilities in terms of alternatives to points to represent dictionary concepts. We leave this for future work.

Conclusions and future work

Concept dictionaries have served as one of the major tools for theory-driven text analysis, producing impressive results across a wide range of problems and tasks. But, in spite of these successes, challenges remain. First, the difficulty of developing and validating broad-coverage sets of words has meant that not all teams have had the skills and resources to build dictionaries in their domains. Second, existing techniques have struggled to apply these dictionaries to shorter texts (such as social media posts).

At the same time, while statistical natural language processing has been going through a period of explosive growth, many of its tools have not been adopted by social scientists due to their atheoretic nature. For many social scientific applications, better classifier performance doesn't help if it can't be related to an underlying model.

In this paper, we introduce DDR, a method designed to bridge the gap between these approaches. Data driven yet conceptually seeded, DDR incorporates the strengths of statistical methods with the theory-driven structure of conceptual dictionaries. By generating a distributed concept representation based on the words in a dictionary, it provides a continuous measure of similarity between a concept and any other word, phrase, or document. This provides a range of benefits for both dictionary authorship and application.

Here, we would like to emphasize that we do not view DDR as a replacement for word count methods. Whenever the object of inquiry is a closed set of words (that is, a fixed set of terms which completely cover a category), word count methods remain more appropriate. For example, many linguistic categories such as pronouns, articles, and conjunctions are composed of a relatively fixed set of terms. The study of these closed classes has proven to be extremely rich for both linguistic and psychological research (Pennebaker, 2011). The notion of similarity on which DDR is based makes little sense in these contexts.

For open class terms, the situation is a little more complicated. For short texts, the decision is clear in favor of DDR. There simply isn't enough context there for word count to extract a clear signal from most dictionaries. For longer documents, the choice is slightly more complicated. While DDR is once again effective in this case, word count's applicability depends on the dictionary structure. If the dictionary includes terms which are sufficiently high frequency in the chosen domain, word count may still be used. In general, though, it is safe to use word counts with closed class terms (or in domains where the relevant words can be completely enumerated) and DDR with open class dictionaries, especially when dealing with shorter documents. DDR doesn't replace existing methods, but rather augments them with a new set of tools.

In terms of dictionary authorship, DDR helps with several major challenges. First, it allows authors to focus on the conceptual core of a category rather than attempting to determine all possible words which might be associated with that category. Given the breadth and dynamic nature of language, a complete enumeration will generally be infeasible. However, with DDR, authors can focus on the key elements that define a category, making use of semantic rather than morphological similarity to find related terms.

Second, DDR allows for limited domain adaptation through the choice of distributed representation. This opens up the potential for a given concept to be more easily explored across a range of domains and application settings. Further, it opens the potential for researchers to make use of representations trained on potentially smaller quantit

ies of domain-specific text, allowing for even more focused adaptation.

Third, DDR provides a means for dictionary authors to directly explore the structure of the dictionaries they are creating. By looking at the relations of candidate dictionary words in the context of a range of distributed representations, dictionary authors have another tool for evaluating and validating their work. The combination of DDR and measurements of semantic coherence allow them to rapidly evaluate the impact of changes to their dictionaries and how the words they have selected hold together. None of this is a replacement for existing psycholinguistic or behavioral validations, however the results shown here suggest that DDR can be a valuable addition to researchers' toolkits.

Further, DDR offers new scope and application for existing dictionaries. In both Study 1 and Study 2, we demonstrated how DDR improved the performance of well validated dictionaries on real world applications. We believe this blend of making existing dictionaries more useful while greatly easing the task of generating new dictionaries to be a powerful combination.

Study 2 in particular demonstrated some of these advantages of DDR. One facet of this comes from DDR returning a smoother measure of similarity between texts and dictionaries. Few of the social media posts contained *any* of the words in the original dictionaries, and so would have been beyond the scope of traditional word-count analysis. However, by applying those dictionaries through DDR, it is possible to generate a smooth measure of similarity between posts and dictionaries, even when there is no word-level overlap between the two. As an increasing amount of social interaction is captured by precisely these sorts of short texts (whether on Twitter, Facebook, or any of the various chat applications), this capability will be increasingly valuable.

Study 1 points to the ability of DDR to provide a simplified version of domain adaptation. By applying a simplified dictionary to a distributed representation trained on that domain, we were able to get better results than using combinations of both larger dictionaries and more extensive distributed representations trained on generic domains. While this is by no means a complete solution to these challenges, it at least provides a small step and the tools made possible with these measures of distributional similarity may help in analyzing these challenges going forward.

A large number of avenues remain for future work. While we have shown a number of intriguing results in terms of particular combinations of dictionaries and representations, we are far from establishing a general rule for which representation will be most appropriate. In fact, the comparison of these strengths and weaknesses for particular conceptual domains may prove to yield a useful window on the underlying structures of those representations.

Additionally, while we have made use of the simplifying assumption that the structure of a concept can be approximated by distance to a single point in semantic space, there is room for further exploration. It remains for future work to determine how more complex models of conceptual structure could provide better mechanisms for evaluation and application.

Nonetheless, DDR provides a new level of flexibility and applicability for theory-driven text analysis. Combining distributed representations and dictionaries, this method makes it possible to leverage the strengths of both. Critically, it does so in a way that takes advantage of existing work. DDR doesn't obsolete current dictionaries, rather it improves their performance and expands their applicability. It doesn't attempt to restructure distributed representations, but rather leverages their strengths to explore theory-driven constructs. In providing a bridge between these two approaches, we hope that it will serve to enrich both.

Acknowledgments This work has been funded in part by NSF IBSS #1520031. Correspondence concerning this article should be addressed to Morteza Dehghani, mdehghan@usc.edu, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061.

References

Box, G. E., & Hill, W. J. (1967). Discrimination among mechanistic models Discrimination among mechanistic models. *Technometrics*, *9*(1), 57–71.

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, 0956797614566658.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429.

Carnap, R. (1959). Logical positivism. New York: The Free Press.

Chen, Q., Li, W., Lei, Y., Liu, X., & He, Y. (2015). Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *ACL (1)* (pp. 419–429).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale: L. Erlbaum.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., ... & Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*, *145*(3), 366.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., et al. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Elsevier.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307.

Frimer, J. A., & Brandt, M. J. (2015). Conservatives display greater happiness but only when they are in power: A linguistic analysis of the U.S. Congress.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. *ICWSM*, *7*(21), 219–222.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.

Gunn, J. F., & Lester, D. (2015). Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. *Suicidologi*, *17*(3).

Haidt, J., Graham, J., & Joseph, C. (2009). Above and below left–right: Ideological narratives and moral foundations. *Psychological Inquiry*, *20*(2–3), 110–119.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Hosmer, D. W. Jr., & Lemeshow, S. (2004). *Applied logistic regression.* Wiley.

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(02), 265–290.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 0261927X13502654.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology*, 263–286.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., & Seligman, M. E. (2014). The online social self an open vocabulary approach to personality. *Assessment*, *21*(2), 158–169.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, *11*, 538–541.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Levy, O., & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations *Proceedings of the eighteenth conference on computational natural language learning*. Association for computational linguistics. Baltimore.

Li, G., & Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, *38*(2), 127–139.

Li, J., Jurafsky, D., & Hovy, E. (2015). When are tree structures necessary for deep learning of representations? arXiv:1503.00185.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 986–1005.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, *2*, 627–666.

Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, *38*(2), 207–221.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128).

Medin, D. L., Bennis, W., & Chandler, M. (2010). Culture and the home-field disadvantage. *Perspectives on Psychological Science*, *5*(6), 708–713.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69.

Menke, J., & Martinez, T. R. (2004). Using permutations instead of student's distribution for p-values in paired-difference algorithm comparisons. In *2004 IEEE international joint conference on neural networks* (Vol. 2, pp. 1331–1335).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mikolov, T., Yih, W. t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL Hlt-naacl* (pp. 746–751).

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Acl* (pp. 236–244).

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M.(2013). The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, *8*(5), e64417.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. *Urbana: Univer. of Illinois Press*, *195*, 36.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 115–124).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*(3), 162–166.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42–45.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP*, *12*, 1532–1543.

Powers, D. M. (1998). Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151–160).

Powers, D. M. W. (2011). Evaluation: from precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*, 37–63.

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *Icwsm*.

Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1988). *Parallel distributed processing* (Vol. 1). IEEE.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268.

Smith, D. A., Rydberg-Cox, J. A., & Crane, G. R. (2000). The perseus project: A digital library for the humanities. *Literary and Linguistic Computing*, *15*(1), 15–25.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation *Proceedings of the Sixteenth ACM conference on conference on information and knowledge management* (pp. 623–632). New York: ACM.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 151–161).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (emnlp)* (Vol. 1631, pp. 1642).

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science*, *8*(1), 113–116.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv:1503.00075.

Tausczik, Y. R., & 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.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178–185.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424).

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Watson, D., & Clark, L. A. (1999). The panas-x: Manual for the positive and negative affect schedule-expanded form.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, *18*(1), 45–55.